ISSN 0141-8130
INTERNATIONAL JOURNAL OF
Biological Macromolecules
STRUCTURE, FUNCTION AND INTERACTIONS
ScienceDirect

Implications derived from S-protein variants of SARS-CoV-2 from six continents

Sk. Sarif Hassan, Kenneth Lundstrom, Debmalya Barh, Raner Jośe Santana Silva, Bruno Silva Andrade, Vasco Azevedo, Pabitra Pal Choudhury, Giorgio Palu, Bruce D. Uhal, Ramesh Kandimalla, Murat Seyran, Amos Lal, Samendra P. Sherchan, Gajendra Kumar Azad, Alaa A.A. Aljabali, Adam M. Brufsky, Ángel Serrano-Aroca, Parise Adadi, Tarek Mohamed Abd El-Aziz, Elrashdy M. Redwan, Kazuo Takayama, Nima Rezaei, Murtaza Tambuwala, Vladimir N. Uversky

PII: S0141-8130(21)02005-5

DOI: https://doi.org/10.1016/j.ijbiomac.2021.09.080

Reference: BIOMAC 19379

To appear in: *International Journal of Biological Macromolecules*

Received date: 15 August 2021

Revised date: 13 September 2021

Accepted date: 13 September 2021

Please cite this article as: S.S. Hassan, K. Lundstrom, D. Barh, et al., Implications derived from S-protein variants of SARS-CoV-2 from six continents, *International Journal of Biological Macromolecules* (2018), https://doi.org/10.1016/j.ijbiomac.2021.09.080

Implications Derived from S-Protein Variants of SARS-CoV-2 from Six Continents

Sk. Sarif Hassan[a,*], Kenneth Lundstrom[b], Debmalya Barh[c], Raner José Santana Silva[d], Bruno Silva Andrade[e], Vasco Azevedo[f], Pabitra Pal Choudhury[g], Giorgio Palu[h], Bruce D. Uhal[i], Ramesh Kandimalla[j], Murat Seyran[k], Amos Lal[l], Samendra P. Sherchan[m], Gajendra Kumar Azad[n], Alaa A. A. Aljabali[o], Adam M. Brufsky[p], Ángel Serrano-Aroca[q], Parise Adadi[r], Tarek Mohamed Abd El-Aziz[s], Elrashdy M. Redwan[t], Kazuo Takayama[u], Nima Rezaei[v], Murtaza Tambuwala[w], and Vladimir N. Uversky[x,*]

[a]Department of Mathematics, Pingla Thana Mahavidyalaya, Maligram, Paschim Medinipur, 721140, West Bengal, India

[b]PanTherapeutics, Rte de Lavaux 49, CH1095 Lutry, Switzerland

[c]Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur, WB, India, & Departamento de Genética, Ecologia e Evolucao, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

[d]Departamento de Ciencias Biologicas (DCB), Programa de Pos Graduacao em Genetica e Biologia Molecular (PPGGBM), UniversidadeEstadual de Santa Cruz (UESC), Rodovia Ilheus-Itabuna, km 16, 45662-900, Ilheus, BA, Brazil

[e]Laboratório de Bioinformática e Química Computacional, Departamento de Ciências Biológicas, Universidade Estadual do Sudoeste da Bahia (UESB), Jequié 45206-190, Brazil

[f]Laboratório de Genética Celular e Molecular, Departamento de Genetica, Ecologia e Evolucao, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, CEP 31270-901, Brazil

[g]Applied Statistics Unit, Indian Statistical Institute, 203 B T Road, Kolkata, 700108, India

[h]Department of Molecular Medicine, University of Padova, Via Gabelli 63, 35121, Padova, Italy[i]Department of Physiology, Michigan State University, East Lansing, MI 48824, USA

[j] Applied Biology, CSIR-Indian Institute of Chemical Technology, Uppal Road, Tarnaka, Hyderabad, 500007, Department of Biocemistry, Kakatiya Medical College, Warangal, Telangana, India

[k]Doctoral Studies in Natural and Technical Sciences (SPL 44), University of Vienna, Währinger Straße, A-1090 Vienna, Austria

[l]Division of Pulmonary and Critical Care Medicine, Mayo Clinic, Rochester, Minnesota, USA [m]Department of Environmental Health Sciences, Tulane University, New Orleans, LA, 70112, USA[n]Department of Zoology, Patna University, Patna, Bihar, India

[o]Department of Pharmaceutics and Pharmaceutical Technology, Yarmouk University, Faculty of Pharmacy, Irbid 566, Jordon

[p]University of Pittsburgh School of Medicine, Department of Medicine, Division of Hematology/Oncology, UPMC Hillman Cancer Center, Pittsburgh, PA, USA

[q] Biomaterials and Bioengineering Lab, Centro de Investigación Traslacional San Alberto Magno, Universidad Católica de Valencia San Vicente Mártir, c/Guillem de Castro, 94, 46001 Valencia, Valencia, Spain

[r]Department of Food Science, University of Otago, Dunedin 9054, New Zealand

[s]Department of Cellular and Integrative Physiology, University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Dr, SanAntonio, TX 78229-3900, USA, & Zoology Department, Faculty of Science, Minia University, El-Minia 61519, Egypt

[t]Faculty of Science, Department of Biological Science, King Abdulazizi University, Jeddah 21589, Saudi Arabia, & Therapeutic and Protective Proteins Laboratory, Protein Research Department, Genetic Engineering and Biotechnology Research Institute, City for Scientific Research andTechnology Applications, New Borg El-Arab, Alexandria 21934, Egypt

[u]Center for iPS Cell Research and Application (CiRA), Kyoto University, Kyoto 606-8507, Japan

[v]Research Center for Immunodeficiencies, Pediatrics Center of Excellence, Children's

Medical Center, Tehran University of Medical Sciences,Tehran, Iran, & Network of Immunity in Infection, Malignancy and Autoimmunity (NIIMA), Universal Scientific Education and Research Network (USERN), Stockholm, Rio de Janeiro, Sweden
[w]*School of Pharmacy and Pharmaceutical Science, Ulster University, Coleraine BT52 1SA, Northern Ireland, UK, UK*
[x]*Department of Molecular Medicine, Morsani College of Medicine, University of South Florida, Tampa, FL 33612, USA*

*Corresponding author
Email addresses: sarimif@gmail.com (Sk. Sarif Hassan), lundstromkenneth@gmail.com (Kenneth Lundstrom), dr.barh@gmail.com (Debmalya Barh), ranerbio@yahoo.com.br (Raner Jos´e Santana Silva), bandrade@uesb.edu.br (Bruno Silva Andrade), vascoariston@gmail.com (Vasco Azevedo), pabitrapalchoudhury@gmail.com (Pabitra Pal Choudhury), giorgio.palu@unipd.it (Giorgio Palu), bduhal@gmail.com (Bruce D. Uhal), ramesh.kandimalla@gmail.com (Ramesh Kandimalla), muratseyran@gmail.com (Murat Seyran), manavamos@gmail.com (Amos Lal), sshercha@tulane.edu (Samendra P. Sherchan), gkazad@patnauniversity.ac.in (Gajendra Kumar Azad), alaaj@yu.edu.jo (Alaa A. A. Aljabali), brufskyam@upmc.edu (Adam M. Brufsky), angel.serrano@ucv.es (Á ngel Serrano-Aroca), pariseadadi@gmail.com (Parise Adadi), mohamedt1@uthscsa.edu (Tarek Mohamed Abd El-Aziz), lradwan@kau.edu.sa (Elrashdy M. Redwan), kazuo.takayama@cira.kyoto-u.ac.jp (Kazuo Takayama), rezaei_nima@tums.ac.ir (Nima Rezaei), m.tambuwala@ulster.ac.uk (Murtaza Tambuwala), vuversky@usf.edu (Vladimir N.Uversky)

Abstract

Spike (S) protein is a critical determinant of the infectivity and antigenicity of SARS-CoV-2. Several mutations in the spike protein of SARS-CoV-2 have already been detected, and their effect in immune system evasion and enhanced transmission as a cause of increased morbidity and mortality are being investigated. From pathogenic and epidemiological perspectives, spike proteins are of prime interest to researchers. This study focused on the unique variants of S proteins from six continents: Asia, Africa, Europe, Oceania, South America, and North America. In comparison to the other five continents, Africa had the highest percentage of unique S proteins (29.1%). The phylogenetic relationship implies that unique S proteins from North America are significantly different from those of the other five continents. They are most likely to spread to the other geographic locations through international travel or naturally by emerging mutations. It is suggested that restriction of international travel should be considered, and massive vaccination as an utmost measure to combat the spread of COVID-19 pandemic. It is also further suggested that the efficacy of existing vaccines and future vaccine development must be reviewed with careful scrutiny, and if needed, further re-engineered based on requirements dictated by new emerging S protein variants.

Keywords: SARS-CoV-2, Invariant residues, Mutations, Spike protein, Continents, Vaccines.

1. Introduction

The world is experiencing a health emergency due to Coronavirus disease (COVID-19), caused by an enveloped positive-sense single-stranded virus, severe acute respiratory syndrome coronavirus (SARS-CoV-2) [1, 2, 3, 4, 5, 6]. The spike (S) protein is a homotrimer present on the surface of the SARS-CoV-2 and recognizes the human host cell surface receptor angiotensin-converting enzyme-2 (ACE2) [7, 8, 9, 10]. The interaction between the S protein of SARS-CoV-2 and its cellular receptor ACE2 is driven by high affinity/avidity. Therefore, neutralization by antibodies does not only require specifically binding antibodies, but antibodies that have high affinity/avidity towards S1 subunit of S protein [11]. It is worth mentioning that this particular aspect is directly related to the variability of S1 (and its isoelectric points) as this may modulate the affinity of binding [12]. The importance of antibody avidity for protection towards SARS-CoV-2 (and other viruses) has been recently reviewed [12]. From the beginning of the second wave of COVID-19 infection, various SARS-CoV-2 variants emerged raising concern of enhanced transmission and mortality of the virus and reduced efficacy of vaccine protection [13, 14]. Some of the studies opposed the perception of SARS-CoV-2 mutations as distinctive pathogenic variants and increased rate of transmissibility were questioned [15, 16]. However, the frequency of the mutant strains within the SARS-CoV-2 population carrying the D614G mutation in the spike protein clearly plays a role in enabling the virus to spread more effectively and rapidly [17]. Epidemiologists have been constantly monitoring the evolution of SARS-CoV-2 with a particular focus on the spike protein and other interacting proteins of the virus [17, 18]. The D614G mutation in the S protein discovered in early 2020 makes the virus able to spread more effectively and rapidly [19]. The D614G mutation has been found to be related with high viral loads in infected patients, and high rate of infections, but not with increased disease severity [20]. Various mutations in the S protein make the SARS-CoV-2 more complex and hence it is more difficult to characterize its severity, infectivity and efficacy of vaccines designed to target S protein. Not all mutations are advantageous to the virus but several mutations or a set of mutations may increase the transmission potential through an increase in receptor binding or the ability to evade the host immune response by altering the surface structures recognized by antibodies [21, 22, 23].

To contain the spread of the COVID-19, it is definitely of high interest to detect and identify various unique emerging variants of S proteins. Additionally, it is also worth investigating the impact of new S protein variants on viral infectivity and potential to spread rapidly as well as to ascertain the origin of the spread of the new variants concerning spike protein variabilities. Accordingly, it might be possible to segregate the set of new variants with respect to individual characteristics of SARS-CoV-2, which would undoubtedly help policy makers to form various strategies to contain the spread of the virus. There are a large number of different SARS-CoV-2 S protein mutant sequences currently available in the NCBI virus database. In this study, all available S protein sequences from six continents Asia, Africa, Europe, North America, South America, and Oceania were analyzed for their uniqueness and variability. An inter-linkage was made among the unique S proteins available on the six continents.

2. Data acquisition and methods

S protein sequences from all six continents (Asia, Africa, Europe, Oceania, South America, and North America) were downloaded in FASTA format from the National Center for Biotechnology Information (NCBI) database (http://www.ncbi.nlm.nih.gov/). Further, FASTA files were processed in *Matlab-2021a* for extracting unique S protein sequences for each continent.

2.1. *Phylogenetic Analysis*

To filter sequences with low quality (unknown amino acids ’X’) and remove redundant sequences, the SeqKit tool was used, with the tools fx2tab and rmdup respectively [24]. The filter removed all sequences that had one or more ’X’ and all redundant sequences (100% identical). The amino acid sequences were aligned using the MegaX program with MUSCLE algorithm, and after it was performed a phylogeny calculation with the Neighbor-joining method, considering 3919 taxa sequences and 530 sites [25, 26]. The alignment was used as input in Archeopteryx 0.9914 with the multiple alignment inference option, following the parameters of maximum allowed gaps ratio 0.5, minimum allowed non-gap sequence length 50 and distance calculator Kimura correction [27]. The phylogenetic trees were analyzed and edited in the Archeopteryx 0.9914tool.

2.2. *Frequency probability of amino acids*

Any protein sequence is composed of twenty different amino acids with various frequencies starting from zero. The ability of occurrence of each amino acid Ai is determined by the formula $\frac{f(A_i)}{l}$where *f(A* $f(A_i)$ denotes the frequency of occurrence of the amino acid A_i in a primary sequence, and *l* stands as the length of an S protein [28]. Hence for each S protein, a twenty-dimensional vector considering the frequency probability of twenty amino acids can be obtained. Based on this frequency probability, the dominance of amino acid density in a given protein is illuminated.

2.3. *Evaluation of normalized amino acid compositions*

The variability of the amino acid compositions of the unique S-proteins from each continent was evaluated using the web-based tool Composition Profiler (*http://www.cprofiler.org/*) that automates detection of enrichment or depletion patterns of individual amino acids or groups of amino acids in query proteins [29]. In this analysis, we used sets of unique S-proteins from each continent as query samples and the amino acid of the original S-protein (UniProt ID: P0DTC2) as a reference sample that provides the background amino acid distribution. Composition profiler generates a bar chart composed of twenty data points (one for each amino acid), where bar heights indicate normalized enrichment or depletion of a given residue. The normalized enrichment/depletion is calculated as

$$\frac{C_{continent} - C_{original}}{C_{original}}$$

where $C_{continent}$ is the content of given residue in the query set of S-proteins in a given continent and $C_{original}$ is the content of the same residue in the original S-protein. For comparison, we generated composition profile of disordered proteins, where normalized composition was evaluated as $\frac{C_{DisProt} - C_{PDB}}{C_{PDB}}$ ($C_{DisProt}$ is content of a given amino acid in the set of intrinsically isordered proteins in the DisProt database [30]; C_{PDB} is content of the given residue in the dataset of fully ordered proteins, PDB-Select-25 [29]). In these analyses, the positive and negative values produced in the compositional profiler indicated enrichment or depletion of the indicated residue, respectively.

2.4. *Amino acid conservation Shannon entropy*

How conserved/disordered the amino acids are organized over S protein is addressed by the information-theoretic measure known as '*Shannon entropy*' (SE). For each S protein, Shannon entropy of amino acid conservation over the amino acid sequence of S protein is computed using the following formula [31, 32]:

For a given amino acid sequence of length *l*, the conservation of amino acids is calculated as follows:

$$SE = -\sum_{i=1}^{20} p_{s_i} \log_{20}(p_{s_i})$$

where $p_{s_i} = \frac{k_i}{l}$; k_i represents the number of occurrences of an amino acid s_i in the given sequence [33].

2.5. *Isoelectric point of a protein sequence*

The isoelectric point (pI), is the pH at which a molecule carries no net electrical charge or is electrically neutral in the statistical mean. We calculate the theoretical pI by using the pKa's of amino acids and summing the net charge across the protein at a given pH (default is typical intracellular pH 7.2), searching with our algorithm for the pH at which the net charge is zero [34]. The isoelectric point is a powerful tool to predict and understand interactions between proteins, proteins and membranes or to determine the presence of protein isoforms [35]. Furthermore, it is noted that the isoelectric point is one of the prime keys for understanding a variety of biochemical properties of protein sequences [35, 36]. Note that the isoelectric point of a protein sequence was computed here using the standard routine of *Matlab-2021a*. This parameter was deployed to characterize the unique S protein sequences, quantitatively.

2.6. *Intrinsic disorder analysis*

Intrinsic disorder predisposition of S protein from the original (Wuhan) version of SARS-CoV-2 was analyzed by a set of six commonly used disorder predictors, such as PONDR® VLXT, PONDR® VL3, PONDR® VSL2B, PONDR® FIT, IUPred2 (Short) and IUPred2 (Long), which were selected for their specific features. The outputs of the evaluation of the per-residue disorder propensity by these tools are represented as real numbers between 1 (ideal prediction of disorder) and 0 (ideal prediction of order) [37, 38, 39, 40, 41]. Thresholds of ≥ 0.15 and ≥ 0.5 were used to identify flexible and disordered residues and regions. Intrinsic disorder profile of this protein was generated by DiSpi/RIDAO web-crawler that combines the outputs of PONDR® VLXT, PONDR® VL3, PONDR® VLS2B, PONDR® FIT, IUPred2 (Short) and IUPred2 (Long) on the one plot and complement them by the errors evaluated for mean disorder profile calculated by averaging profiles of individual predictors. Analysis of intrinsic disorder predisposition of unique variants of S protein was conducted by PONDR® VSL2B. This tool is commonly used in the analysis of disorder predisposition of proteins and systematically shows good performance in various comparative analyses, including the recently conducted Critical assessment of protein intrinsic disorder prediction (CAID) experiment, where PONDR® VLS2B was recognized as predictor #3 of the 43 evaluated methods [42].

3. Results

We first determined the set of unique S protein sequences from each continent. Further, every unique S protein from a continent was compared with other unique S proteins from five continents, and the lists of the same are presented in Tables 12-17. Also, the variability of the S proteins from each continent was shown using Shannon entropy and isoelectric point.

3.1. *Unique spike proteins in the continents*

In Table 1, the number of total sequences, unique sequences and percentages are presented. Note that, a complete list of unique S protein accessions and their names (continent-wise) were made available in *supplementary file-1*. Note that, sequence accession is renamed as *Ck* where *C* stands for continent code (Asia:AS, Africa:AF, Oceania:O, Europe:U, South America:SA, and North America:NA), and *k* denotes the serial number.

Table 1: Percentages of continent-wise unique spike (S) proteins

Continent	**Total S proteins (T)**	**Unique S proteins (U)**	**Percentage, continent-wise**	**Percentage, worldwide**
			$\frac{U}{T}\times 100$	$\frac{U}{16143}\times 100$
Africa	984	286	29.065	1.772
Asia	2314	432	18.669	2.676
Europe	1006	187	18.588	1.158
Oceania	9920	1121	11.300	6.944
South	464	71	15.302	0.440

America				
North America	113072	14046	12.422	87.010
Worldwide	127760	16143	12.635	—

The highest percentage (29.065%) of unique S proteins were found in Africa though the total number of available sequences is significantly low as compared with that from other continents. Almost similar amounts (in percentage) of unique S sequence variations were found in Asia and Europe. Among the total 127760 S proteins embedded in SARS-CoV-2 genomes, only 16143 (12%) unique S proteins were detected so far, and notably most of the unique variants (87%) were found in NorthAmerica only.

For each continent, the unique spike (S) proteins were matched with other unique proteins from the rest of the five continents, and a total number of such identical pairs are presented accordingly in the matrix (Table 2).

Table 2: The total continent-wise number of identical S proteins

Continent-wise	**Asia**	**Africa**	**Europe**	**North America**	**Oceania**	**South America**
Asia	–	25	27	169	17	17
Africa	25	–	15	71	13	5
Europe	27	15	–	76	9	8
North America	169	71	76	–	49	31
Oceania	17	13	9	49	–	5
South America	17	5	8	31	5	–
Total continent-wise	255	129	135	396	93	66
Unique residue S proteins	177	157	52	13650	1028	5

From Table 2, it was observed that, in each continent there is still a significant percentage of unique spike variations available, which are not shared with any rest of the continents. Such percentages of unique variations of S proteins in Asia, Africa, Europe, Oceania, South America, and North America were 41%, 55%, 28%, 92%, 7%, and 97% respectively. The lists of pairs of identical S proteins of SARS-CoV-2 originating from six continents are presented in Tables 9-11 (See *Appendix*). The lists of unique S proteins (from a particular continent), which were found to be identical with some unique spike proteins from other five continents, are presented in Tables (12-17) (See *Appendix*).

The frequency and percentage of invariant residue positions, where no amino acid change was detected so far in the uniqueS proteins available in each continent, are presented in Table 3.

Table 3: The total number and percentage of invariant residue positions among 1273 positions in unique S proteins

	Africa	**Asia**	**Europe**	**Oceania**	**South America**	**North America**
Total Freq.	902	695	948	731	1070	89
Percentage	70.86	54.60	74.47	57.42	84.05	6.99

Frequency of invariant residue positions in unique S proteins from each continent

The highest number of mutations (lowest number of invariant residue position, 6.99%) (Table 3) were detected in the unique S proteins from North America where 12.42% unique S protein sequences were present as mentioned in Table 1. Likewise, the lowest number (15.95%) of mutations in unique S proteins were observed in South America where 15.3% unique S sequences were found. Only 29.14% residues of 1273 in the unique S proteins were mutated, although a significantly higher number (29.065%) of unique sequences were found in Africa among the other five continents. The unique S proteins from Europe possessed only 25.5% mutations, whereas 45.5% mutations were detected in the unique S proteins from Asia although the same percentage (18.5%) of unique spike proteins were found (Tables 1 and 3). Further it was observed that 11.3% of the unique S proteins from Oceania possessed 42.58% mutations.

3.2. *Phylogenetic relationship among unique S-protein variants*

We collected 204440 Spike protein sequences from (NCBI and GAISED databases). Upon filtering, 191536 redundant sequences were removed and 12904 unique sequences (corresponding to 6.31% of the initial number of sequences) were selected for phylogenetic analysis.

The resultant phylogeny for unique amino acid sequences from SARS-CoV-2 S-protein, revealed a tree with polyphyletic groups, as well as showing sequences from different countries grouping together in the same clade (See *Supplementary Figure 1*). On the other hand, after the Archaepteryx analysis it was identified five predominant sequence groups between different spike variants from different countries (Figure 1).

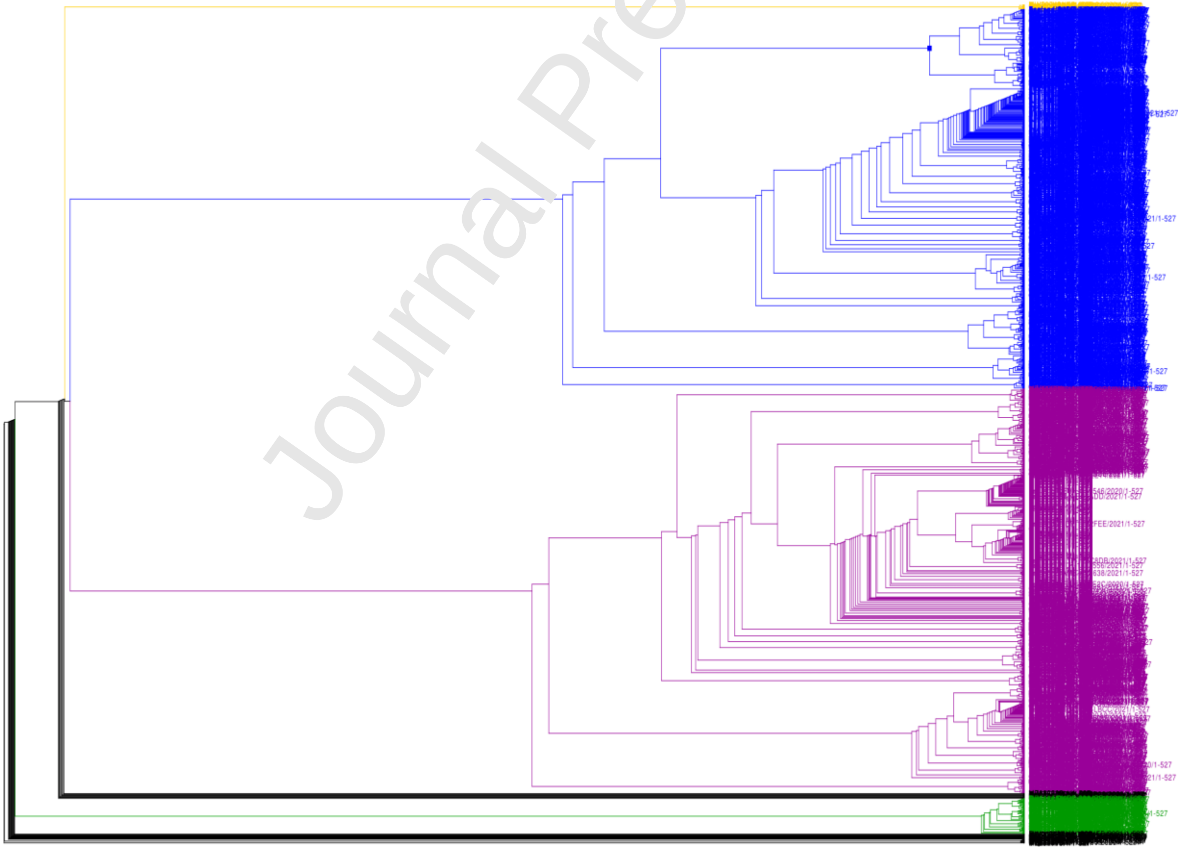


Figure 1: SARS-CoV-2 spike amino acid phylogeny after group clustering. After Archeopteryx analysis it can be identified five groups: yellow, blue, magenta, green and black.

In this case, it can be verified that the different group colors are formed by sequences from the same continents, but it was identified that grouped sequences from different continent together. Then, after these analyses it could be possible to assume that we have at least five unique SARS-CoV-2 spike variants indicating possibilities for new ways for most specific vaccination and drug development.

3.3. *Variability through normalized amino acid composition*

Additional information on the variability of the amino compositions of the unique S-proteins from each continent relative to the composition of original S-protein from Wuhan was retrieved using the web-based tool Composition Profiler (*http://www.cprofiler.org/*). Results of this analysis are shown in Figure 2A, which clearly shows the presence of some noticeable amino acid composition variability among unique S-proteins from different continents. Since individual S proteins are different from each other and from the original S-protein mostly in very limited number of residues, the range of changes in the normalized enrichment/depletion of a given residue is rather limited (compare scales of Y axis in Figures 2A and 2B, where a composition profile of the intrinsically disordered proteins is shown for comparison).

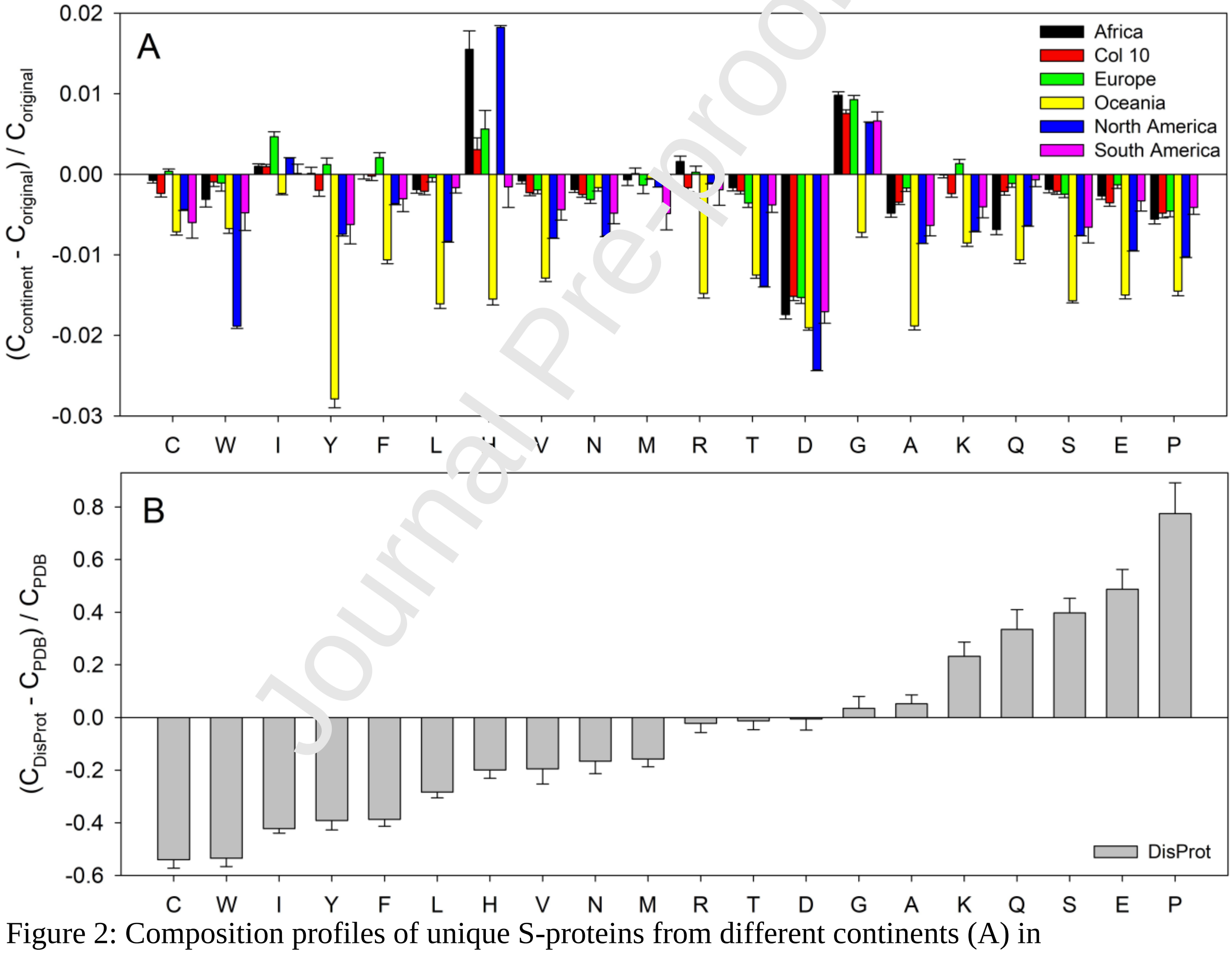


Figure 2: Composition profiles of unique S-proteins from different continents (A) in comparison with the composition profile of typical intrinsically disordered protein (B).

On an average, unique S-proteins form Oceania were found to have the most variability in terms of normalized amino acid composition. This was followed by the unique S-proteins from North America. Curiously, Figure 2A shows that although the normalized content of individual residues in the unique S-proteins from Oceania is always below that of the original S-protein, S-proteins from other continents might have relative excess of some residues. For example, some unique S-proteins from almost all continents can be enriched in glycine or histidine residues, whereas some European S-proteins can also be relatively enriched in cysteine, isoleucine, tyrosine, phenylalanine, and lysine residues (see positive green bars in Figure 2A). Another interesting observation is that the different sets of S-proteins are typically characterized by rather noticeable variability of the normalized content of most residues. The noticeable exception is given by aspartate, depletion in which is almost uniform between all the unique S-proteins from all the continents.

3.4. Variability through intrinsic disorder analysis

Next, we looked at the correlation between the frequencies of the mutations in unique S proteins from different locations and intrinsic disorder predisposition of this protein. Figure 3A shows distribution of frequency of mutations within the amino acid sequence of this protein. It is seen that almost each residue of has at least one mutation in different variants currently found in the globe. In fact, only 15 residues (Met$_1$, Leu$_{996}$, Ile$_{997}$, Gly$_{999}$, Leu$_{1001}$, Tyr$_1$007, Val$_1$008, Gln$_{1010}$, Ile$_{1013}$, Arg$_{1019}$, His$_{1049}$, Gln$_{1054}$, Thr$_{1105}$, Asn$_{1119}$, and Leu$_{1270}$) of the 1273-residue long S protein were never mutated as of the time of this analysis. Curiously, nine of these never-changed residues are concentrated within the short region (residues 996-1019). Figure 3A also shows that mutation frequencies are unevenly distributed within the amino acid sequence of S protein and that the region (residues 675-691) surrounding the furin cleavage site (residues 680-686) seems to be characterized by high mutation frequency. In fact, although the average per-residue frequency of mutations of the entire protein is equal to 4.6, the mutation frequency of the 675-691 region is two-fold higher (9.2). Comparison of the mutation frequency profile (Figure 3A) with the per-residue intrinsic disorder predisposition profile generated for the original (Wuhan) version of S-protein by a set of commonly used disorder predictors (Figure 3B) indicates that there is some weak correlation between these two parameters, with regions showing more disorder typically undergoing more frequent mutations. Again, Figure 3B shows that region containing furin cleavage site is among the most disordered segments of the S protein (if not the most disordered one).

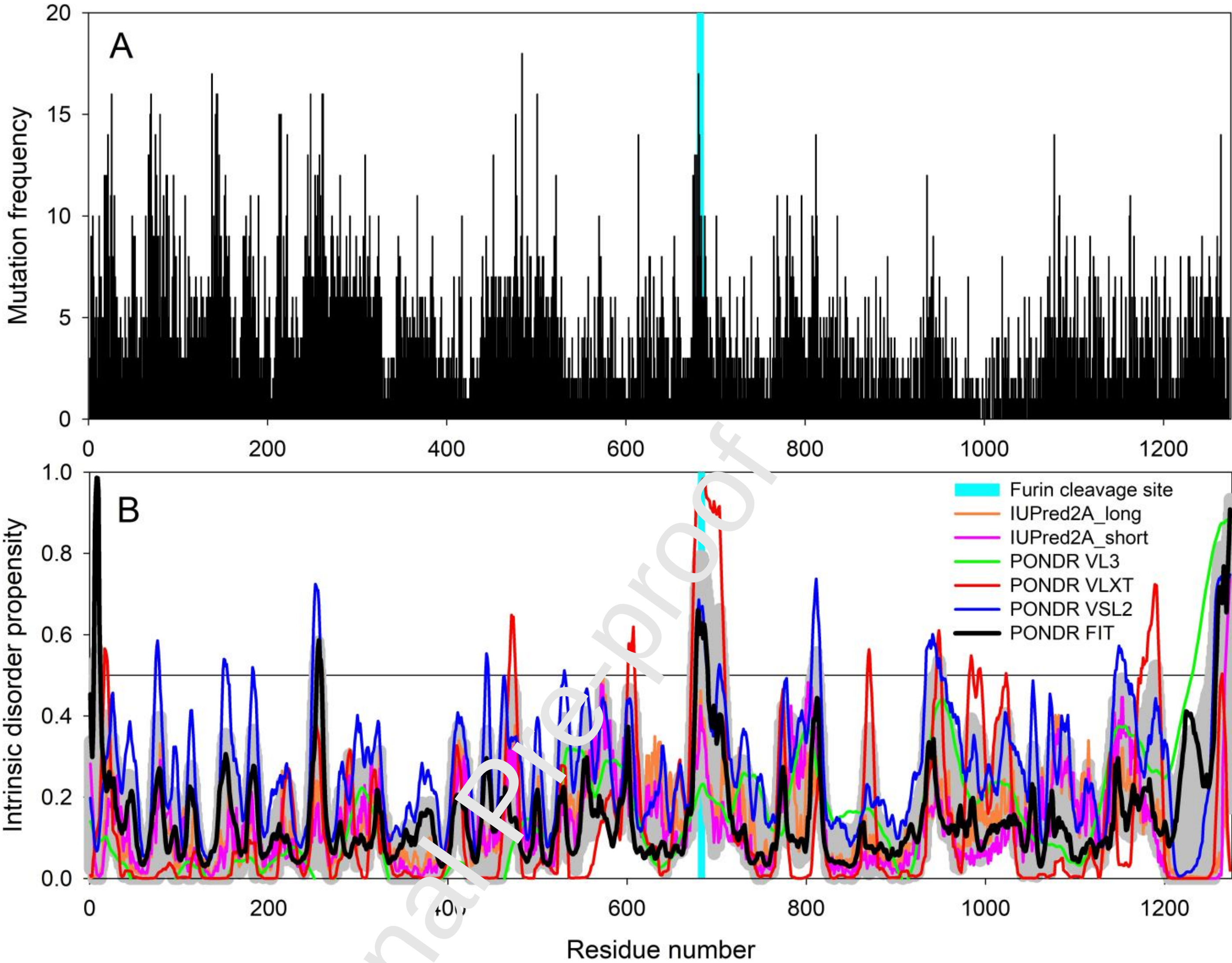


Figure 3: Correlation of the sequence variability of unique variants of S protein with the intrinsic disorder predisposition of this protein. A. Frequencies of mutations observed at each residue of S protein in various locations. B. Intrinsic disorder predisposition of the original (Wuhan) version of S protein analyzed by a set of commonly used disorder predictors. In both plots, position of the furin cleavage site is shown as a cyan vertical bar.

Figure 4 provides further quantification of the per-residue mutability and disorder predisposition of S protein. Here, dependencies of the mutation frequencies on the corresponding disorder score evaluated by PONDR® VSL2 are shown for six geo-locations. In Africa, S protein has 902, 296, 60, 11, and 4 residues with 0, 1, 2, 3, and 4 mutations, which are characterized by the mean disorder scores of 0.27 ± 0.15, 0.28 ± 0.16, 0.30 ± 0.18, 0.29 ± 0.18, and 0.40 ± 0.19, respectively. In Asia, 694, 437, 111, 27, 3, and 1 residues of S protein with 0, 1, 2, 3, 4, and 5 mutations are characterized by the mean disorder scores of 0.26 ± 0.14, 0.28 ± 0.15, 0.32 ± 0.16, 0.37 ± 0.18, 0.52 ± 0.22, and 0.19, respectively. In Europe, 948, 265, 55, and 5 residues of S protein with 0, 1, 2, and 3 mutations have the mean disorder scores of 0.27 ± 0.15, 0.28 ± 0.15, 0.33 ± 0.18 and 0.35 ± 0.27, respectively. In Europe, 948, 265, 55, and 5 residues of S protein with 0, 1, 2, and 3 mutations have the mean disorder scores of 0.27 ± 0.15, 0.28 ± 0.15, 0.33 ± 0.18, and 0.35 ± 0.27, respectively. Oceania's S protein has 722, 427, 107, 13, and 4 residues with 0, 1, 2, 3, and 4 mutations, which are showing the disorder scores of 0.26 ± 0.14, 0.28 ± 0.16, 0.30 ± 0.17, 0.35 ± 0.21, and 0.27 ± 0.15, respectively. In S proteins from South America variants, 1070, 193, and 10 residues with 0, 1, and 2 mutations have mean disorder scores of 0.27 ± 0.15, 0.36 ± 0.13, and 0.23 ± 0.10, respectively. Finally, in North America, S protein underwent most mutations and has 23, 351, 323, 234, 167, 108, 42, 12, 6, 4, 1, 1, and 1 residues with 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, and 13 mutations characterized by the mean disorder scores of 0.28 ± 0.17, 0.25 ± 0.13, 0.25 ± 0.13, 0.29 ± 0.17, 0.31 ± 0.19, 0.38 ± 0.16, 0.4 ± 0.17, 0.22 ± 0.16, 0.42 ± 0.22, 0.18, 0.25, and 0.45, respectively.

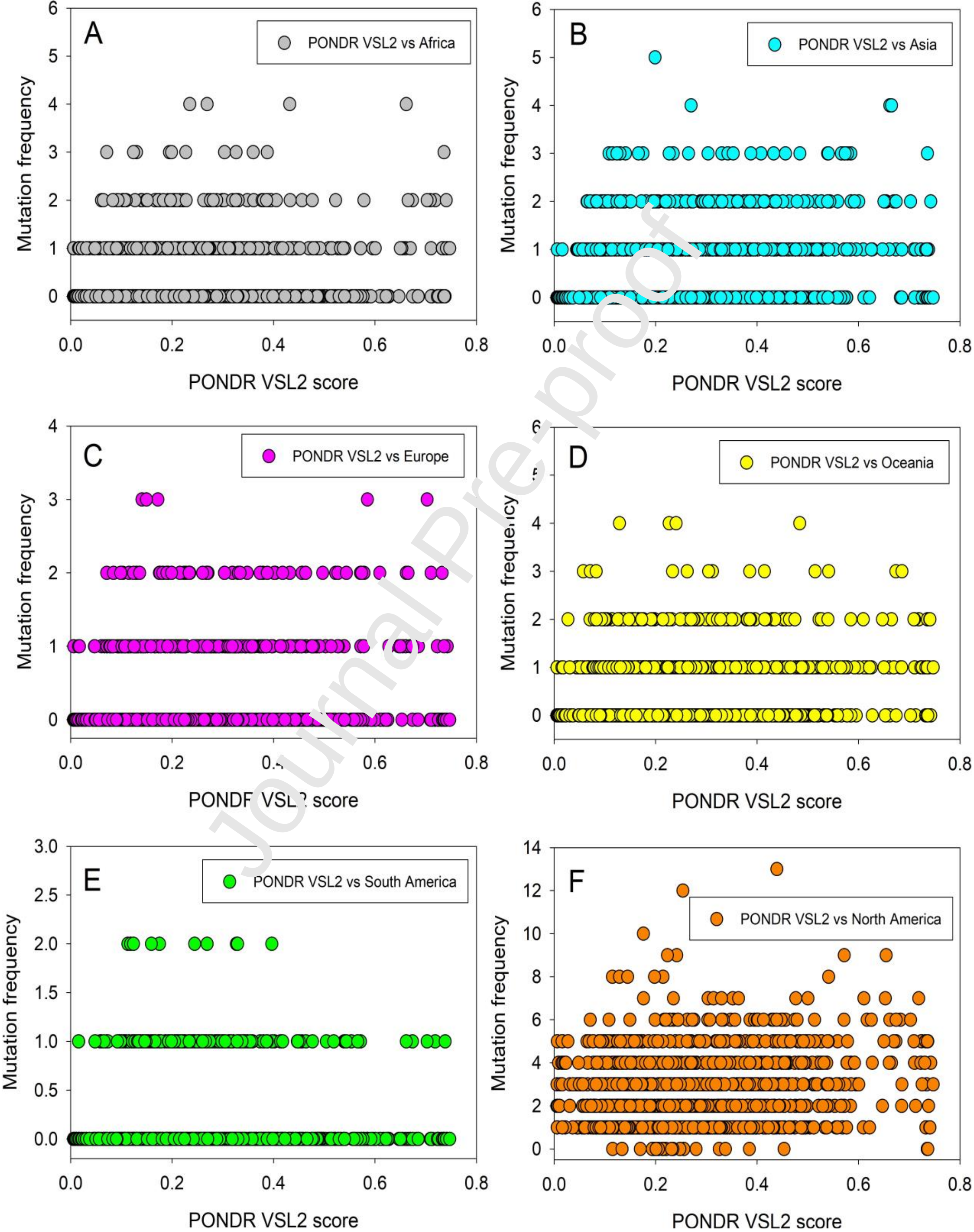


Figure 4: Correlation of frequency of amino acid substitutions at a given residue of S-protein and the corresponding intrinsic disorder score of this residues within the sequence of the

original (Wuhan) version of S protein. Individual plots reflect distributions within sequences of variants found in different locations: A. Africa; B. Asia; C. Europe, D. Oceania, E. South America; F. North America.

The most frequently mutated residue is Tyr_{248} (13 mutations) followed by Val_{213} and Thr_{108} with 12 and 10 mutations, respectively, all three from unique S protein variants found in North America. This analysis shows that here is a general trend, where residues with higher disorder levels are mutated more frequently.

3.5. *Variability of unique spike proteins*

We quantitatively determined the variations in the unique S proteins on six continents. The variations were captured through the frequency distribution of amino acids present, Shannon entropy (amount of conservation of amino acids in a given sequence), and molecular weights and isoelectric points of a given protein sequence.

3.5.1. *Variations in the frequency distribution of amino acids*

The frequency of each amino acid was computed for each unique S protein available in six continents (*Supplementary file-2*). Maximum and minimum frequencies of amino acids present in the unique S proteins from different continents are presented in Table 4.

Table 4: Maximum and minimum frequencies of amino acids present in the unique spike proteins from different continents

Max and Min of Frequencies		**A**	**R**	**N**	**D**	**C**	**Q**	**E**	**G**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**
Africa	*Max*	80	44	89	62	41	63	49	84	19	79	109	62	15	78	60	101	98	13	56	98
	Min	73	40	85	58	38	59	45	78	14	73	102	57	13	72	55	94	90	11	49	93
Asia	*Max*	80	44	89	63	41	63	49	84	19	78	110	62	15	79	59	101	101	13	57	98
	Min	73	39	80	55	36	56	45	76	15	72	100	55	13	68	52	90	90	11	49	90
Europe	*Max*	80	43	89	63	41	63	49	84	19	79	110	62	15	79	59	101	98	13	57	99
	Min	75	38	84	59	39	59	46	79	16	74	102	58	13	74	54	96	90	11	50	93
Oceania	*Max*	81	43	90	62	41	63	49	84	18	78	109	62	15	79	59	100	98	12	56	99
	Min	72	3	8	5	3	5	4	7	1	7	97	5	1	7	5	92	8	1	4	8

	in		7	1	8	6	7	4	4	5	1		6	3	1	2		8	0	3	9
Sout	*M*	82	4	9	6	4	6	4	8	2	7	11	6	1	8	6	10	9	1	5	1
h	*ax*		4	1	3	2	4	9	5	0	9	1	4	5	0	0	2	9	3	8	0
Ame																					0
rica	*M*	60	3	6	4	3	3	3	6	1	5	82	4	9	5	4	76	7	8	3	8
	in		2	3	6	2	9	4	3	1	5		3		5	3		7		6	2
Nort	*M*	80	4	8	6	4	6	4	8	1	7	10	6	1	7	5	10	9	1	5	9
h	*ax*		3	9	2	1	3	8	3	8	8	9	2	4	9	8	1	8	2	7	8
Ame	*M*	75	3	8	5	3	5	4	7	1	7	10	5	1	7	5	92	9	1	5	9
rica	*in*		8	2	7	7	9	5	9	6	3	5	7	3	3	7		3	1	0	2

All S protein sequences are leucine (L) and serine (S) rich. Tryptophan (W) and methionine (M) were presented with the least frequencies (Table 4). The widest variation in frequency distributions of the twenty amino acids over the unique S proteins was found in North America.

To obtain quantitative variations in the unique S proteins available in each continent, differences between maximum and minimum vectors (20 dimensions) were obtained (Table 5), and then Euclidean distances between the difference vectors was calculated (Table 6).

Table 5: Matrix presenting the difference between maximum and minimum frequencies of amino acids present in the unique S proteins on each continent

Difference matrix	**A**	**R**	**N**	**D**	**C**	**Q**	**E**	**G**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**
Africa	7	4	4	4	3	4	4	6	5	6	7	5	2	6	5	7	8	2	7	5
Asia	7	5	9	8	5	[illegible]	4	8	4	6	10	7	2	11	7	11	11	2	8	8
Europe	5	5	5	4	2	4	3	5	3	5	8	4	2	5	5	5	8	2	7	6
Oceania	9	6	9	4	5	6	5	10	3	7	12	6	2	8	7	8	10	2	13	10
South America	5	5	7	5	4	4	3	4	2	5	4	5	1	6	1	9	5	1	7	6
North America	22	12	28	17	10	25	15	22	9	24	29	21	6	25	17	26	22	5	22	18

Table 6: Pairwise Euclidean distances among the difference vectors of each continent

Distance matrix	**Africa**	**Asia**	**Europe**	**Oceania**	**South America**	**North America**
Africa	0.00	11.70	4.69	12.77	8.49	66.80
Asia	11.70	0.00	13.00	9.06	14.04	57.02
Europe	4.69	13.00	0.00	13.30	8.49	68.38
Oceania	12.77	9.06	13.30	0.00	16.03	56.84
South America	8.49	14.04	8.49	16.03	0.00	69.02
North America	66.80	57.02	68.38	56.84	69.02	0.00

Based on the distance matrix, a phylogenetic relationship was derived among the continents (Figure 2).

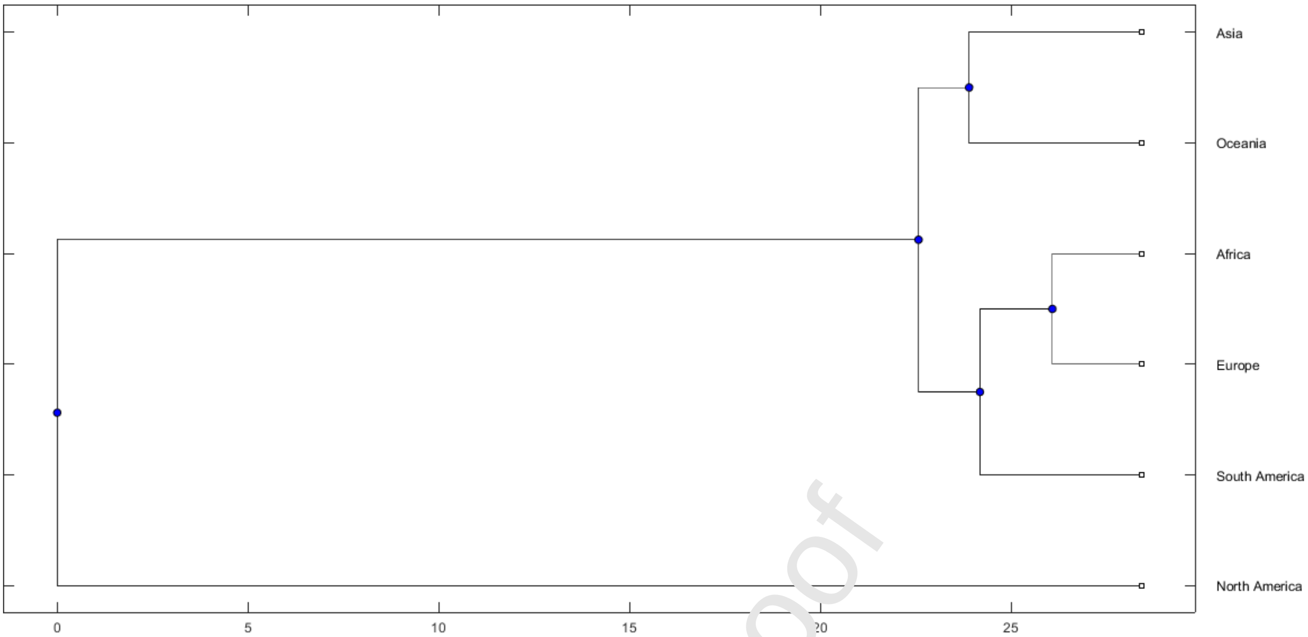


Figure 5: Phylogenetic relationship among the six continents based on the variability of unique spike proteins available in each continent.

Variations based on the frequency distribution of amino acids present in the S proteins make North America (which belongs to the rightmost branch of the tree) distant from the other five continents (Figure 2). Variations among the unique spike proteins from Asia and Oceania turned out to be similar, and they belong to the same level of leaves of the far left branch of the tree. Africa and Europe were found to be the closest in terms of variations based on the frequency distribution of amino acids over the unique spike proteins from each continent. Variability of spike proteins from South America has distant resemblance to that of Africa/Europe as estimated in the phylogeny. The frequencies of amino acid distribution in each unique S protein from each continent are presented in Figures 6 and 7 (See *Appendix*). The widest variations of the frequency distribution of amino acids present in S proteins were observed in North America as wide band was observed in Figure 7. Individual frequency distributions of amino acids in Asia and Oceania seem very close as it was observed from the phylogeny (Figure 5).

3.5.2. Variability through Shannon entropy

In principle, for a random amino acid sequence, the Shannon entropy (SE) is one. Here Shannon entropy for each S protein sequence was computed using the formula stated in section 2.2 (*Supplementary file-2*). It was found that the highest and lowest SEs of S proteins from all continents were 0.9643 and 0.9594 respectively. That is, the length of the largest interval is 0.005 which is sufficiently small. Also note that the length of the smallest interval was 0.001 which occurred in the SEs of S proteins from South America. Within this realm, the widest variation of SEs was noticed among the unique S proteins of North America. All other four intervals (considering lowest and highest) of SEs of all the unique S proteins from four continents Africa, Asia, Oceania and Europe were contained in the interval of North America and contain that of South America.

Table 7: Interval of Shannon entropy of unique S proteins from six different continents

SE: Continent	Interval of SEs
SE of S protein: Africa	(0.960825, 0.963239)
SE of S protein: Asia	(0.961471, 0.963326)

SE of S protein: Europe	(0.961539, 0.963254)
SE of S protein: North America	(0.95934, 0.964314)
SE of S protein: Oceania	(0.961525, 0.963042)
SE of S protein: South America	(0.961589, 0.962895)

Among all (20^{1273}) possible amino acids (20 in number) sequences of length 1273, Nature(?) had selected only a fraction to make S proteins of SARS-CoV-2, and interestingly SEs of them were kept within a very small interval. From the SEs which were close to 1, the S protein sequences are expected to be pseudo-random. Variation of SEs for all unique S proteins from each continent is shown in Figures 5 and 6 (See *Appendix*). Conservation of amino acids present over each S protein from each continent is different from one another which is depicted by the zig-zag nature of SEs plots (Figure 8 and 9).

3.5.3. *Variability through isoelectric point*

For each S protein sequence from each continent isoelectric point (pI) was computed (Supplementary file-3). Intervals (considering minimum and maximum) pIs of unique spike proteins from each continent were tabulated in Table 8.

Table 8: Interval of isoelectric point of unique S proteins from six different continents

pI: Continent	**Interval of PIs**
pI of S protein: Africa	(6.44, 7.09)
pI of S protein: Asia	(6.21, 7.08)
pI of S protein: Europe	(6.21, 6.99)
pI of S protein: North America	(5.61, 7.79)
pI of S protein: Oceania	(6.31, 7.09)
pI of S protein: South America	(6.36, 6.99)

It was noticed that pIs for all the unique S proteins from the six continents were distributed between 5.61 and 7.79. The largest interval of pIs was found for the unique S proteins from North America. Therefore, the widest varieties of unique S proteins were found in North America.

The degree of non-linearity of the plots of pIs for each protein from each continent shows wide variations of unique S proteins (Figures 10 and 11 (See *Appendix*)).

4. Discussion and concluding remarks

Various mutations in S proteins lead to the evolution of new variants of SARS-CoV-2 [43]. Naturally, our attention was captured to characterize unique S protein variants which were embedded in SARS-CoV-2 genomes infecting millions people worldwide [44]. As of May 7, 2021, there are 127760 patients infected with SARS-CoV-2 with 16143 S protein variants, which undoubtedly well-organized by means of amino acids composition and conservation as it was depicted by Shannon entropy and isoelectric point. Among the unique spike proteins present in a continent, many of them are common in other continents as well (Table 2). On the other hand, there are still a handful of unique spike protein variants residing in each continent. Considering the nature and biological implications of the new variants of SARS-CoV-2 caused by different mutations in S proteins, the appearance of several unique S variants in SARS-CoV-2 is certainly a worrying event. [45]. There are still many unique S protein variants in all continents that may spread from person to person through close communities or by spontaneous mutations causing a condition that may become alarming.

Comparative analysis revealed the presence of some weak correlation between the per-residue mutability and intrinsic disorder predisposition of S protein, with residues possessing higher disorder predisposition typically showing higher mutation rates as well. For example, the mean disorder score of 89 residues that were mutated 10-18 times is 0.35 ± 0.18 as compared to the mean disorder score of 0.26 ± 0.12 for 155 residues with 0 and 1 mutations. Curiously, the most disorder region of the S protein (residues 675-691), which includes the furin cleavage site (residues 680-686), was shown to be characterized by high mutation frequency, with Pro_{681} (which was mutated 17 times) being second most frequently mutated residue of this protein.

We observed that unique S proteins from North America have mutations in almost every amino acid residue position (1184 out of 1273), while unique spike variants from the other continents only have mutations in 16 to 20% of residues. So, even if international travel is limited, S proteins from these five continents will likely acquire mutations at other residue positions where mutations have already been found in the specific variants from North America due to natural evolution. Based on the amino acid frequency distributions in the S protein variants from all the continents, a phylogenetic relationship among the continents was drawn. The phylogenetic relationship implies that unique S proteins from North America were found to be significantly different from that of other five continents. Therefore, the possibility of spreading the unique variants originated from North America to the other geographic locations by means of international travel is high, and numerous mutations have been detected already in the unique variants from North America. Of note, South America infection/herd immunity status may have summarized by Manaus city example (the capital of Amazonas state in northern Brazil) where by June 2020 to October 2020 SARS-CoV-2 prevalence among Manaus population increased from ~60% to ~70%, a condition which may mirror acquisition of herd immunity [46]. By January 2021 Manaus had a huge resurgence in cases due to emergence of a new variant known as P.1, which was responsible for nearly 100% of the new case [47]. Although the population may have then reached a high herd immunity threshold, there is still a risk of resurgence of new immunity-escape variants, which raises important questions. For example, 1. Is post-infection herd immunity not enough for protection and should it be combined with vaccination? 2. Will the crucial viral variants (mutations) be listed by WHO and recommended to be included in "next generation vaccines"? [48, 49]. In addition, we cannot yet exclude the possibility of serious mutations in the viral RBD emerging in India and the USA [48].

Let us have a brief glance at the potential consequences of the mutations in S-protein from the viewpoint of protective immunity towards SARS-CoV-2. It is known that the protective immunity towards infection and disease depends on the presence of high avidity antibodies. This is because high avidity of neutralizing antibodies, which is defined as the strength of antibody-target epitope interaction, plays an important role in antibody-mediated protection against viral infections [12]. High avidity (functional affinity) is established during affinity maturation, as the avidity of IgG is low during acute infection and reaches high values several weeks or months later [50, 51]. Importantly, incomplete avidity maturation of IgG directed towards the often leads to the failure of the protection against viral infections and/or resultant diseases, as was shown for varicella zoster virus (VZV), cytomegalovirus (CMV), the measles virus, Dengue virus, the respiratory syncytial virus (RSV), Simian human immunodeficiency virus (SHIDV) [52, 53, 54, 55, 56, 57, 58, 59]. Since the interaction between the receptor-binding domain of SARS-CoV-2 spike protein and angiotensin-converting enzyme-2 (ACE2) on target cells is characterized by high affinity, it is expected that the protective anti-SARS-CoV-2 antibodies should possess high affinity/avidity to be able to block this high affinity RBD-ACE2 interaction [11]. Recently, it was shown that the serological response to SARS-CoV-2 is frequently characterized by the incomplete maturation of avidity [60, 50]. It was also proposed that such incomplete avidity maturation represents an essential strategy of coronaviruses determining high probability of repeated waves of reinfections with these viruses due to the short-lasting protective immunity [61, 62, 12]. Furthermore, an unexpected scenario was recently uncovered, where the natural SARS-CoV-2 infection does not lead to the establishment of a high avidity immune response and therefore does not have a good chance for the development of complete protection against SARS-CoV-2 and for establishment of herd immunity [63]. On the contrary, complete avidity maturation was achieved with two rounds of vaccination, whereas the quality of the immune response after natural infection was similar to that generated by one vaccination step and did not reach the quality of complete vaccination with two steps. Therefore, the scenario occurring in Manaus city can be considered on the basis of these new findings. In fact, it is obvious now that despite the high COVID-19 prevalence reached in this city, no herd immunity could be expected retrospectively, as natural infection is insufficient for the establishment of a high avidity immune response and related development of complete protection against SARS-CoV-2 [63]. Therefore, it seems likely that the herd immunity can only be reached through at least two vaccination steps [63]. It is also expected that the herd immunity might be partially or completely overrun by the novel SARS-CoV-2 variants showing higher affinity of RBD-ACE2 interaction that that of the originally infecting variant.

Hence in the near future, we can expect to experience more new SARS-CoV-2 variants which might cause third, fourth, and fifth etc. waves of COVID-19. Therefore, massive vaccination is necessary to combat COVID-19, and of course, existing vaccines must be reviewed, and if needed further re-engineered may be required based on newly emerging S protein variants.

Altogether, data presented in this study indicate that although unique variants of the SARS-CoV-2 S protein are rather abundant, they are unevenly distributed among continents, with Africa possessing highest percentage of unique S variants, and with unique S proteins found in North America being noticeably different from the variants seeing on other continents. It is likely that these unique variants can spread to continents where they have not been detected before. Furthermore, this inhomogeneity raises an important question on why the currently observed difference in the number of unique variants of S protein (reflecting frequency of its mutagenesis) is so great. It cannot be easily explained by the differences between the continents in the number of COVID-19 patients (reported SARS-CoV-2-positive cases). In fact, according to Worldometer, as of September 10, 2021, there were 8,087,058, 72,407,564, 56,618,705, 181,742, 50,106,216, and 37,213,429 recorded COVID-19 cases in Africa, Asia, Europe, Oceania, North America, and South America, respectively. Obviously, these infection levels do not correlate with the corresponding numbers of unique S protein variants (see Tables 1 and 2). There is also no strong correlation between the reported S protein variability and levels of genomic sequencing in different continents (which serves now as a real-time molecular/genomic SARS-CoV-2 surveillance). In fact, it was reported recently that as of 5 July 2021, 25,284 whole-genome sequences from Africa (0.32% of all reported SARS-CoV-2-positive cases from that continent), 146,562 from Asia (0.30% coverage), 1,292,415 from Europe (2.35% coverage), 692,704 from North America (1.75% coverage), 37,913 from South America (0.12% coverage) and 20,613 from Oceania (25% coverage) had been generated [64]. Therefore, although these numbers reflecting levels of the continent-wise coverage show a heavy bias toward the regions and countries with more specialized genomic facilities, programs, and research projects, there is no strong correlation between the coverage and established S-protein variability [65, 66].

An intriguing possible mechanism of the observed differences in the rates of virus evolution is the presence of a conceivable variability of the ACE2 gene in different continents that might have an impact on the variability of the viral protein as well. In line with this idea, it was recently reported that the expression levels of ACE2 can be elevated up to 50% due to the differences in the frequency of the rs2285666 polymorphism (the TT-plus strand or AA-minus strand alternate allele) among Europeans and Asians, with this difference playing a significant role in the SARS-CoV-2 susceptibility [67, 68]. Similarly, based on the comprehensive analyses of the allelic frequencies of the polymorphisms in the ACE2, TMPRSS2, TMPRSS11A, cathepsin L (CTSL), and elastase (ELANE) genes in populations from the American, African, European, and Asian continents it was concluded that the non-coding sequences of these genes encoding proteins related to the SARS-CoV-2 cell entry contain numerous polymorphisms with possible functional consequences [69].

CRediT author statement

Sk. Sarif Hassan: Study design, Formal analysis, Investigation, Methodology, Visualization, Writing - Original Draft,Writing - Review & Editing, Administration;
Kenneth Lundstrom: Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing;
Debmalya Barh: Formal analysis, Investigation, Methodology, Visualization, Writing - Review & Editing;
Raner José Santana Silva: Formal analysis, Investigation, Methodology, Visualization, Writing - Review & Editing; **Bruno Silva Andrade**: Formal analysis, Investigation, Methodology, Visualization, Writing - Review & Editing; **Vasco Azevedo**: Formal analysis, Investigation, Methodology, Visualization, Writing - Review & Editing;
Pabitra Pal Choudhury: Formal analysis, Writing - Review & Editing;
Giorgio Palu: Formal analysis, Writing - Review & Editing;
Bruce D. Uhal: Formal analysis, Writing - Review & Editing;
Ramesh Kandimalla: Formal analysis, Writing - Review & Editing;
Murat Seyran: Formal analysis, Writing - Review & Editing;
Amos Lal: Formal analysis, Writing - Review & Editing;
Samendra P. Sherchan: Formal analysis, Writing - Review & Editing;
Gajendra Kumar Azad: Formal analysis, Writing - Review & Editing;
Alaa A. A. Aljabali: Formal analysis, Writing - Review & Editing;
Adam M. Brufsky: Formal analysis, Writing - Review & Editing;
Angel Serrano-Aroca: Formal analysis, Writing - Review & Editing;
Parise Adadi: Formal analysis, Writing - Review & Editing;
Tarek Mohamed Abd El-Aziz: Formal analysis, Writing - Review & Editing;
Elrashdy M. Redwan: Formal analysis, Writing - Review & Editing;
Kazuo Takayama: Formal analysis, Writing - Review & Editing;
Nima Rezaei: Formal analysis, Writing - Review & Editing;
Murtaza Tambuwala: Formal analysis, Writing - Review & Editing;
Vladimir N. Uversky: Study design, Formal analysis, Methodology, Investigation, Visualization, Writing - Original Draft,Writing - Review & Editing;
All authors read and approved the final version of the manuscript.

Conflict of interests statement

Authors have no conflict of interest to declare.

References

[1] S. M. Lokman, M. Rasheduzzaman, A. Salauddin, R. Barua, A. Y. Tanzina, M. H. Rumi, M. I. Hossain, A. Z. Sid- diki, A. Mannan, M. M. Hasan, Exploring the genomic and proteomic variations of sars-cov-2 spike glycoprotein: a computational biology approach, Infection, Genetics and Evolution 84 (2020) 104389.

[2] Á. Serrano-Aroca, K. Takayama, A. Tuñón-Molina, M. Seyran, S. S. Hassan, P. P.

Choudhury, V. N. Uversky, K. Lund- strom, P. Adadi, G. Palù, et al., Carbon-based nanomaterials: Promising antiviral agents to combat covid-19 in the microbial resistant era, ACS NanoPMID: 33826850. doi:10.1021/acsnano.1c00629.

[3] S. Hassan, S. Ghosh, D. Attrish, P. P. Choudhury, A. A. Aljabali, B. D. Uhal, K. Lundstrom, N. Rezaei, V. N. Uversky, M. Seyran, et al., Possible transmission flow of sars-cov-2 based on ace2 features, Molecules 25 (24) (2020) 5906.

[4] M. Martí, A. Tuñón-Molina, F. L. Aachmann, Y. Muramoto, T. Noda, K. Takayama, Á. Serrano-Aroca, Protective face mask filter capable of inactivating sars-cov-2, and methicillin-resistant staphylococcus aureus and staphylococcus epidermidis, Polymers 13 (2) (2021) 207.

[5] S. S. Hassan, D. Attrish, S. Ghosh, P. P. Choudhury, V. N. Uversky, A. A. Aljabali, K. Lundstrom, B. D. Uhal, N. Rezaei, M. Seyran, et al., Notable sequence homology of the orf10 protein introspects the architecture of sars-cov-2, International Journal of Biological Macromolecules 181 (2021) 801–809.

[6] S. S. Hassan, A. A. Aljabali, P. K. Panda, S. Ghosh, D. Attrish, P. P. Choudhury, M. Seyran, D. Pizzol, P. Adadi, T. M. Abd El-Aziz, et al., A unique view of sars-cov-2 through the lens of orf8 protein, Computers in biology and medicine (2021) 104380.

[7] L. Zhang, C. B. Jackson, H. Mou, A. Ojha, H. Peng, B. D. Quinlan, E. S. Rangarajan, A. Pan, A. Vanderheiden, M. S. Suthar, et al., Sars-cov-2 spike-protein d614g mutation increases virion spike density and infectivity, Nature communications 11 (1) (2020) 1–9.

[8] L. Guruprasad, Human sars cov-2 spike protein mutations, Proteins: Structure, Function, and Bioinformatics 89 (5) (2021) 569–576.

[9] R. Henderson, R. J. Edwards, K. Mansouri, K. Janowska, V. Stalls, S. M. Gobeil, M. Kopp, D. Li, R. Parks, A. L. Hsu, et al., Controlling the sars-cov-2 spike glycoprotein conformation, Nature structural & molecular biology 27 (10) (2020) 925–933.

[10] M. Seyran, K. Takayama, V. N. Uversky, K. Lundstrom, G. Palù, S. P. Sherchan, D. Attrish, N. Rezaei, A. A. Aljabali, S. Ghosh, et al., The structural basis of accelerated host cell entry by sars-cov-2, The FEBS journal (2020).

[11] I. Khatri, F. J. Staal, J. J. Van Dongen, Blocking of the high-affinity interaction-synapse between sars-cov-2 spike and human ace2 proteins likely requires multiple high-affinity antibodies: an immune perspective, Frontiers in immunology 11 (2020).

[12] G. Bauer, The potential significance of high avidity immunoglobulin g (igg) for protective immunity towards SARS-CoV-2, International Journal of Infectious Diseases 106 (2021) 61–64.

[13] E. B. Hodcroft, D. B. Domman, D. J. Snyder, K. Oguntuyo, M. Van Diest, K. H. Densmore, K. C. Schwalm, J. Femling, J. L. Carroll, R. S. Scott, et al., Emergence in late 2020 of multiple lineages of sars-cov-2 spike protein variants affecting amino acid position 677, MedRxiv (2021).

[14] Z. Ke, J. Oton, K. Qu, M. Cortese, V. Zila, L. McKeane, T. Nakane, J. Zivanov, C. J. Neufeldt, B. Cerikan, et al., Structures and distributions of sars-cov-2 spike proteins on intact virions, Nature 588 (7838) (2020) 498–502.

[15] O. A. MacLean, R. J. Orton, J. B. Singer, D. L. Robertson, No evidence for distinct types in the evolution of sars-cov-2, Virus Evolution 6 (1) (2020) veaa034.

[16] L. van Dorp, M. Acman, D. Richard, L. P. Shaw, C. E. Ford, L. Ormond, C. J. Owen, J. Pang, C. C. Tan, F. A. Boshier, et al., Emergence of genomic diversity and recurrent mutations in sars-cov-2, Infection, Genetics and Evolution 83 (2020) 104351.

[17] J. Zhang, Y. Cai, T. Xiao, J. Lu, H. Peng, S. M. Sterling, R. M. Walsh, S. Rits-Volloch, H. Zhu, A. N. Woosley, et al., Structural impact on sars-cov-2 spike protein by d614g substitution, Science 372 (6541) (2021) 525–530.

[18] S. E. Park, Epidemiology, virology, and clinical features of severe acute respiratory syndrome-coronavirus-2 (sars-cov-2; coronavirus disease-19), Clinical and experimental pediatrics 63 (4) (2020) 119.

[19] E. Callaway, The coronavirus is mutating-does it matter?, Nature 585 (7824) (2020) 174–177.

[20] B. Korber, W. M. Fischer, S. Gnanakaran, H. Yoon, J. Theiler, W. Abfalterer, N. Hengartner, E. E. Giorgi, T. Bhat- tacharya, B. Foley, et al., Tracking changes in sars-cov-2 spike: evidence that d614g increases infectivity of the covid-19virus, Cell 182 (4) (2020) 812–827.

[21] E. Volz, V. Hill, J.T. McCrone, A. Price, D. Jorgensen, Á. O'Toole, J. Southgate, R. Johnson, B. Jackson, F.F. Nascimento, et al., Evaluating the effects of sars-cov-2 spike mutation d614g on transmissibility and pathogenicity, Cell184 (1) (2021) 64–75.

[22] T. C. Williams, W. A. Burgers, Sars-cov-2 evolution and vaccines: cause for concern?, The Lancet Respiratory Medicine9 (4) (2021) 333–335.

[23] H. Tegally, E. Wilkinson, M. Giovanetti, A. Iranzadeh, V. Fonseca, J. Giandhari, D. Doolabh, S. Pillay, E. J. San, N. Msomi, et al., Emergence of a sars-cov-2 variant of concern with mutations in spike glycoprotein., Nature (2021).

[24] W. Shen, S. Le, Y. Li, F. Hu, Seqkit: a cross-platform and ultrafast toolkit for fasta/q file manipulation, PloS one11 (10) (2016) e0163962.

[25] S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, Mega x: molecular evolutionary genetics analysis across computing platforms, Molecular biology and evolution 35 (6) (2018) 1547.

[26] R. C. Edgar, Muscle: multiple sequence alignment with high accuracy and high throughput, Nucleic acids research 32 (5) (2004) 1792–1797.

[27] M. V. Han, C. M. Zmasek, phyloxml: Xml for evolutionary biology and comparative genomics, BMC bioinformatics 10 (1) (2009) 1–6.

[28] D. J. Brooks, J. R. Fresco, A. M. Lesk, M. Singh, Evolution of amino acid frequencies in proteins over deep time: inferred order of introduction of amino acids into the genetic code, Molecular Biology and Evolution 19 (10) (2002) 1645–1655.

[29] V. Vacic, V. N. Uversky, A. K. Dunker, S. Lonardi, Composition profiler: a tool for discovery and visualization of amino acid composition differences, BMC bioinformatics 8 (1) (2007) 1–7.

[30] M. Sickmeier, J. A. Hamilton, T. LeGall, V. Vacic, M. S. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V. N. Uversky, et al., Disprot: the database of disordered proteins, Nucleic acids research 35 (suppl 1) (2007) D786–D793.

[31] S. S. Hassan, D. Attrish, S. Ghosh, P. P. Choudhury, B. Roy, Pathogenetic perspective of missense mutations of orf3a protein of sars-cov-2, Virus Research (2021) 198441.

[32] S. S. Hassan, P. P. Choudhury, B. Roy, S. S. Jana, Missense mutations in sars-cov2 genomes from Indian patients, Genomics 112 (6) (2020) 4622–4627.

[33] B. J. Strait, T. G. Dewey, The shannon information entropy of protein sequences, Biophysical journal 71 (1) (1996) 148–155.

[34] P. G. Righetti, Determination of the isoelectric point of proteins by capillary isoelectric focusing, Journal of chromatography A 1037 (1-2) (2004) 491–499.

[35] F. S. Stekhoven, M. Gorissen, G. Flik, The isoelectric point, a key to understanding a variety of biochemical problems: a minireview, Fish physiology and biochemistry 34 (1) (2008) 1–8.

[36] G. Adair, The chemistry of the proteins and amino acids, Annual review of biochemistry 6 (1) (1937) 163–192.

[37] P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown, A. K. Dunker, Sequence complexity of disordered protein, Proteins: Structure, Function, and Bioinformatics 42 (1) (2001) 38–48.

[38] K. Peng, P. Radivojac, S. Vucetic, A. K. Dunker, Z. Obradovic, Length-dependent prediction of protein intrinsic disorder, BMC bioinformatics 7 (1) (2006) 1–17.

[39] K. Peng, S. Vucetic, P. Radivojac, C. J. Brown, A. K. Dunker, Z. Obradovic, Optimizing long intrinsic disorder predictors with protein evolutionary information, Journal of bioinformatics and computational biology 3 (01) (2005) 35–60.

[40] B. Xue, R. L. Dunbrack, R. W. Williams, A. K. Dunker, V. N. Uversky, Pondr-fit: a meta-predictor of intrinsically disordered amino acids, Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics 1804 (4) (2010) 996–1010.

[41] B. Mészáros, G. Erdős, Z. Dosztányi, Iupred2a: context-dependent prediction of protein disorder as a function of redox state and protein binding, Nucleic acids research 46 (W1) (2018) W329–W337.

[42] M. Necci, D. Piovesan, S. C. Tosatto, Critical assessment of protein intrinsic disorder prediction, Nature methods 18 (5)(2021) 472–481.

[43] A. Baum, B. O. Fulton, E. Wloga, R. Copin, K. E. Pascal, V. Russo, S. Giordano, K. Lanza, N. Negron, M. Ni, et al., Antibody cocktail to sars-cov-2 spike protein prevents rapid mutational escape seen with individual antibodies, Science 369 (6506) (2020) 1014–1018.

[44] Z. Liu, L. A. VanBlargan, L.-M. Bloyet, P. W. Rothlauf, R. E. Chen, S. Stumpf, H. Zhao, J. M. Errico, E. S. Theel, M. J. Liebeskind, et al., Identification of sars-cov-2 spike mutations that attenuate monoclonal and serum antibody neutralization, Cell host & microbe 29 (3) (2021) 477–488.

[45] B. Dearlove, E. Lewitus, H. Bai, Y. Li, D. B. Reeves, M. G. Joyce, P. T. Scott, M. F. Amare, S. Vasan, N. L. Michael, et al., A sars-cov-2 vaccine candidate would likely match all currently circulating variants, Proceedings of the National Academy of Sciences 117 (38) (2020) 23652–23662.

[46] L. F. Buss, C. A. Prete, C. M. Abrahim, A. Mendrone, T. Salomon, C. de Almeida-Neto, R. F. França, M. C. Belotti, M. P. Carvalho, A. G. Costa, et al., Three-quarters attack rate of sars-cov-2 in the brazilian amazon during a largelyunmitigated epidemic, Science 371 (6526) (2021) 288–292.

[47] C. Aschwanden, Five reasons why covid herd immunity is probably impossible. Nature 591 (7851) (2021) 520–522.

[48] R. K. Gupta, Will sars-cov-2 variants of concern affect the promise of vaccines? Nature Reviews Immunology (2021) 1–2.

[49] E. M. Redwan, Covid-19 pandemic and vaccination build herd immunity, Eur Rev Med Pharmacol Sci 25(2) (2021) 577–579.

[50] G. Bauer, The variability of the serological response to sars-corona virus-2: potential resolution of ambiguity through determination of avidity (functional affinity), Journal of Medical Virology 93 (1) (2021) 311–322.

[51] K. Hedman, M. Lappalainen, M. Söderlund, L. Hedman, Avidity of igg in serodiagnosis of infectious diseases, Reviews in Medical Microbiology 4 (3) (1993) 123–129.

[52] A. K. Junker, P. Tilley, Varicella-zoster virus antibody avidity and igg-subclass patterns in children with recurrent chickenpox, Journal of medical virology 43 (2) (1994) 119–124.

[53] S. B. Boppana, W. J. Britt, Antiviral antibody responses and intrauterine transmission after primary maternal cytomegalovirus infection, Journal of Infectious Diseases 171 (5) (1995) 1115–1121.

[54] T. Lazzarotto, S. Varani, P. Spezzacatena, L. Gabrielli, P. Pradelli, B. Guerra, M. P. Landini, Maternal igg avidity and igm detected by blot as diagnostic tools to identify pregnant women at risk of transmitting cytomegalovirus, Viral immunology 13 (1) (2000) 137–141.

[55] S. Seo, Y. Cho, J. Park, Serologic screening of pregnant Korean women for primary human

cytomegalovirus infection using igg avidity test, The Korean journal of laboratory medicine 29 (6) (2009) 557–562.

[56] M. Kaneko, M. Ohhashi, T. Minematsu, J. Muraoka, K. Kusumoto, H. Sameshima, Maternal immunoglobulin g avidityas a diagnostic tool to identify pregnant women at risk of congenital cytomegalovirus infection, Journal of Infection andChemotherapy 23 (3) (2017) 173–176.

[57] A. Puschnik, L. Lau, E. A. Cromwell, A. Balmaseda, S. Zompi, E. Harris, Correlation between dengue-specific neutralizing antibodies and serum avidity in primary and secondary dengue virus 3 natural infections in humans, PLoS neglected tropical diseases 7 (6) (2013) e2274.

[58] M. F. Delgado, S. Coviello, A. C. Monsalvo, G. A. Melendi, J. Z. Hernandez, J. P. Batalle, L. Diaz, A. Trento, H.-Y. Chang, W. Mitzner, et al., Lack of antibody affinity maturation due to poor toll-like receptor stimulation leads to enhanced respiratory syncytial virus disease, Nature medicine 15 (1) (2009) 34–41.

[59] L. Lai, D. Vödrös, P. A. Kozlowski, D. C. Montefiori, R. L. Wilson, V. L. Akerstrom, L. Chennareddi, T. Yu, S. Kannanganat, L. Ofielu, et al., Gm-csf DNA: an adjuvant for higher avidity Igg, rectal Iga, and increased protection against the acute phase of a shiv-89.6 p challenge by a DNA/mva immunodeficiency virus vaccine, Virology 369 (1) (2007) 153–167.

[60] G. Bauer, F. Struck, P. Schreiner, E. Staschik, E. Soutschek, M. Motz, The serological response to SARS corona virus-2 is characterized by frequent incomplete maturation of functional affinity (avidity) (2020).

[61] A. W. Edridge, J. Kaczorowska, A. C. Hoste, M. Bakker, M. Klein, K. Loens, M. F. Jebbink, A. Matser, C. M. Kinsella, P. Rueda, et al., Seasonal coronavirus protective immunity is short-lasting, Nature medicine 26 (11) (2020) 1691–1693.

[62] M. Galanti, J. Shaman, Direct observation of repeated infections with endemic coronaviruses, The Journal of infectiousdiseases 223 (3) (2021) 409–415.

[63] F. Struck, P. Schreiner, E. Staschik, K. Wochinz-Richter, S. Schulz, E. Soutschek, M. Motz, G. Bauer, Vaccination versusinfection with SARS-CoV-2: Establishment of a high avidity Igg response versus incomplete avidity maturation, Journal ofMedical Virology (2021).

[64] WHO-COVID-19-Dashboard, COVID live update, https://covid19.who.int/table (accessed September 12, 2021).

[65] S. L. Wu, A. N. Mertens, Y. S. Crider, A. Nguyen, N. N. Pokpongkiat, S. Djajadi, A. Seth, M. S. Hsiang, J. M. Colford, A. Reingold, et al., Substantial underestimation of SARS-CoV-2 infection in the United States, Nature communications11 (1) (2020) 1–10.

[66] M. Mohanan, A. Malani, K. Krishnan, A. Acharya, Prevalence of SARS-CoV-2 in Karnataka, India, Jama 325 (10) (2021) 1001–1003.

[67] R. Asselta, E. M. Paraboschi, A. Mantovani, S. Duga, Ace2 and tmprss2 variants and expression as candidates to sex and country differences in COVID-19 severity in Italy, Aging (Albany NY) 12 (11) (2020) 10087.

[68] A. Srivastava, A. Bandopadhyay, D. Das, R. K. Pandey, V. Singh, N. Khanam, N. Srivastava, P. P. Singh, P. K. Dubey, A. Pathak, et al., Genetic association of ace2 rs2285666 polymorphism with COVID-19 spatial distribution in India, Frontiers in genetics 11 (2020) 1163.

[69] G. Vargas-Alarcón, R. Posadas-Sánchez, J. Ramírez-Bello, Variability in genes related to SARS-CoV-2 entry into host cells (ace2, tmprss2, tmprss11a, elane, and ctsl) and its potential use in association studies, Life Sciences 260 (2020) 118313.

Appendix

Table 9: List of pairs of identical spike proteins of SARS-CoV-2 originated from six continents

Spike: Asia-Europe	Spike: Asia-Africa	Spike: Asia-Oceania	Spike: Asia-South America	Spike:Asia-North America
(A14, U2)	(A14, AF2)	(A15, O5)	(A31, SA1)	(A1, NA7)
(A15, U3)	(A15, AF3)	(A77, O43)	(A67, SA4)	(A8, NA231)
(A30, U8)	(A26, AF19)	(A95, O58)	(A148, SA13)	(A12, NA902)
(A31, U9)	(A71, AF48)	(A109, O83)	(A180, SA19)	(A14, NA928)
(A33, U11)	(A93, AF58)	(A128, O201)	(A191, SA22)	(A15, NA992)
(A36, U17)	(A128, AF72)	(A138, O370)	(A200, SA25)	(A19, NA1131)
(A43, U18)	(A138, AF76)	(A142, O373)	(A207, SA27)	(A23, NA1445)
(A69, U23)	(A142, AF79)	(A148, O377)	(A211, SA30)	(A28, NA2065)
(A77, U26)	(A148, AF82)	(A166, O387)	(A213, SA32)	(A30, NA3228)
(A93, U28)	(A161, AF88)	(A206, O388)	(A219, SA33)	(A31, NA3313)
(A95, U30)	(A164, AF92)	(A213, O390)	(A234, SA35)	(A32, NA3438)
(A105, U34)	(A166, AF101)	(A253, O398)	(A280, SA41)	(A33, NA3477)
(A128, U52)	(A191, AF115)	(A277, O400)	(A284, SA42)	(A34, NA3658)
(A134, U54)	(A206, AF118)	(A284, O407)	(A335, SA61)	(A43, NA3752)
(A135, U57)	(A213, AF120)	(A305, O504)	(A340, SA63)	(A44, NA3768)
(A148, U63)	(A275, AF130)	(A359, O1076)	(A373, SA68)	(A58, NA3911)
(A213, U80)	(A276, AF131)	(A404, O1104)	(A404, SA71)	(A69, NA4028)
(A234, U84)	(A277, AF134)			(A71, NA4051)
(A239, U88)	(A279, AF137)			(A76, NA4169)
(A265, U94)	(A282, AF138)			(A77, NA4243)
(A284, U99)	(A292, AF147)			(A78, NA4270)
(A286, U100)	(A379, AF229)			(A85, NA4296)
(A333, U121)	(A394, AF247)			(A89, NA4375)
(A340, U124)	(A404, AF263)			(A90, NA4394)
(A379, U151)	(A430, AF278)			(A91, NA4436)
(A404, U181)				(A93, NA4448)
(A430, U187)				(A95, NA4508)

Spike: Asia-North America	Spike: Asia-North America	Spike: Asia-North America	Spike: Asia-North America	Spike:Asia-North America
(A96, NA4537)	(A166, NA5819)	(A214, NA6445)	(A267, NA6903)	(A345, NA9597)
(A97, NA4541)	(A170, NA5927)	(A215, NA6465)	(A273, NA6916)	(A348, NA9612)
(A100, NA4559)	(A171, NA5977)	(A216, NA6492)	(A274, NA6936)	(A351, NA9663)
(A101, NA4620)	(A173, NA5992)	(A217, NA6499)	(A275, NA6944)	(A354, NA9674)
(A102, NA4637)	(A174, NA6060)	(A218, NA6510)	(A276, NA6949)	(A356, NA9724)
(A103, NA4658)	(A175, NA6067)	(A219, NA6515)	(A277, NA6962)	(A357, NA9763)
(A105, NA4715)	(A177, NA6071)	(A221, NA6527)	(A278, NA6969)	(A358, NA9776)
(A109, NA4861)	(A178, NA6080)	(A222, NA6540)	(A279, NA7000)	(A359, NA9792)
(A111, NA4897)	(A180, NA6101)	(A223, NA6550)	(A280, NA7015)	(A360, NA9834)
(A114, NA5001)	(A181, NA6142)	(A224, NA6553)	(A282, NA7025)	(A367, NA10276)
(A115, NA5022)	(A182, NA6148)	(A230, NA6602)	(A283, NA7056)	(A373, NA10342)
(A121, NA5105)	(A183, NA6155)	(A233, NA6616)	(A284, NA7090)	(A375, NA10442)
(A122, NA5137)	(A191, NA6185)	(A234, NA6622)	(A286, NA7129)	(A378, NA11135)

(A126, NA5151)	(A193, NA6193)	(A235, NA6630)	(A291, NA7198)	(A379, NA11225)
(A127, NA5182)	(A195, NA6244)	(A238, NA6659)	(A292, NA7227)	(A380, NA11305)
(A128, NA5194)	(A196, NA6258)	(A239, NA6661)	(A293, NA7249)	(A381, NA11560)
(A133, NA5471)	(A198, NA6276)	(A244, NA6683)	(A304, NA7576)	(A383, NA11874)
(A134, NA5485)	(A199, NA6293)	(A245, NA6687)	(A322, NA8509)	(A386, NA13280)
(A135, NA5516)	(A200, NA6299)	(A247, NA6707)	(A323, NA8519)	(A387, NA13307)
(A138, NA5538)	(A201, NA6305)	(A249, NA6713)	(A324, NA8565)	(A388, NA13362)
(A140, NA5574)	(A205, NA6324)	(A253, NA6751)	(A325, NA8570)	(A391, NA13404)
(A148, NA5595)	(A206, NA6334)	(A254, NA6756)	(A333, NA9283)	(A394, NA13438)
(A158, NA5644)	(A207, NA6373)	(A255, NA6780)	(A335, NA9324)	(A395, NA13444)
(A159, NA5645)	(A210, NA6388)	(A257, NA6794)	(A341, NA9425)	(A396, NA13465)
(A161, NA5666)	(A211, NA6406)	(A258, NA6810)	(A342, NA9455)	(A399, NA13554)
(A163, NA5722)	(A212, NA6424)	(A264, NA6857)	(A343, NA9568)	(A401, NA13614)
(A164, NA5744)	(A213, NA6429)	(A265, NA6862)	(A344, NA9592)	(A404, NA13635)
				(A405, NA13668)
				(A408, NA13704)
				(A413, NA13841)
				(A418, NA13913)
				(A419, NA13948)
				(A430, NA14000)
				(A431, NA14026)

Table 10: List of pairs of identical spike proteins of SARS-CoV-2 originated from different continents

Spike: Africa-Europe	**Spike: Africa-North America**	**Spike: Africa-North America**	**Spike: Africa-Oceania**	**Spike: Africa-South America**	**Spike: Europe-North America**
(AF2, U2)	(AF2, NA928)	(AF121, NA6566)	(AF1, O3)	(AF82, SA13)	(U2, NA928)
(AF3, U3)	(AF3, NA992)	(AF123, NA6628)	(AF3, O5)	(AF115, SA22)	(U3, NA992)
(AF31, U10)	(AF8, NA1298)	(AF125, NA6816)	(AF71, O148)	(AF117, SA26)	(U4, NA1221)
(AF58, U28)	(AF9, NA1348)	(AF128, NA6848)	(AF72, O201)	(AF120, SA32)	(U7, NA2680)
(AF69, U45)	(AF31, NA3387)	(AF130, NA6944)	(AF76, O370)	(AF263, SA71)	(U8, NA3228)
(AF72, U52)	(AF34, NA3583)	(AF131, NA6949)	(AF79, O373)		(U9, NA3313)
(AF82, U63)	(AF38, NA3797)	(AF133, NA6953)	(AF82, O377)		(U10, NA3387)
(AF120, U80)	(AF46, NA3986)	(AF134, NA6962)	(AF101, O387)		(U11, NA3477)
(AF123, U85)	(AF47, NA3988)	(AF137, NA7000)	(AF118, O388)		(U18, NA3752)
(AF145, U103)	(AF48, NA4051)	(AF138, NA7025)	(AF120, O390)		(U22, NA3895)
(AF195, U119)	(AF50, NA4061)	(AF145, NA7199)	(AF134, O400)		(U23, NA4028)
(AF229, U151)	(AF51, NA4117)	(AF146, NA7224)	(AF179, O751)		(U26, NA4243)
(AF230, U154)	(AF58, NA4448)	(AF147, NA7227)	(AF263, O1104)		(U28, NA4448)
(AF263, U181)	(AF64, NA4832)	(AF149, NA7286)		**Spike: Oceania-South America**	(U30, NA4508)
(AF278, U187)	(AF69, NA5149)	(AF151, NA7299)		(O377, SA13)	(U34, NA4715)
	(AF71, NA5188)	(AF152, NA7300)		(O389, SA28)	(U36, NA4780)

(AF72, NA5194)	(AF154, NA7375)		(O390, SA32)	(U38, NA4837)
(AF73, NA5202)	(AF156, NA7453)		(O402, SA42)	(U41, NA4989)
(AF76, NA5538)	(AF165, NA7553)		(O1104, SA71)	(U42, NA5083)
(AF82, NA5595)	(AF168, NA7644)			(U45, NA5149)
(AF83, NA5606)	(AF179, NA8514)			(U47, NA5167)
(AF88, NA5666)	(AF195, NA9264)			(U52, NA5194)
(AF90, NA5693)	(AF196, NA9265)			(U53, NA5282)
(AF92, NA5744)	(AF223, NA10257)			(U54, NA5485)
(AF99, NA5818)	(AF227, NA10943)			(U55, NA5490)
(AF101, NA5819)	(AF229, NA11225)			(U57, NA5516)
(AF103, NA5829)	(AF230, NA11456)			(U63, NA5595)
(AF104, NA5830)	(AF231, NA11576)			(U66, NA5627)
(AF105, NA5837)	(AF247, NA13438)			(U72, NA6096)
(AF108, NA5874)	(AF248, NA13478)			(U76, NA6240)
(AF114, NA6178)	(AF254, NA13578)			(U78, NA6399)
(AF115, NA6185)	(AF263, NA13635)			(U79, NA6421)
(AF118, NA6334)	(AF268, NA13798)			(U80, NA6429)
(AF119, NA6390)	(AF271, NA13870)			(U82, NA6450)
(AF120, NA6429)	(AF278, NA14000)			(U84, NA6622)
	(AF283, NA14015)			(U85, NA6628)
				(U88, NA6661)
				(U90, NA6704)

Table 11: List of pairs of identical spike proteins of SARS-CoV-2 originated from different continents

Spike: Europe-North America	Spike: Europe-Oceania	Spike: North America-Oceania	Spike: North America-Oceania	Spike: South America-North America
(U92, NA6723)	(U3, O5)	(NA992, O5)	(NA6751, O398)	(NA3313, SA1)
(U93, NA6775)	(U26, O43)	(NA3873, O28)	(NA6962, O400)	(NA4550, SA5)
(U94, NA6862)	(U30, O58)	(NA4024, O36)	(NA7060, O401)	(NA4720, SA7)
(U98, NA7057)	(U52, O201)	(NA4243, O43)	(NA7090, O402)	(NA4989, SA11)
(U99, NA7090)	(U63, O377)	(NA4508, O58)	(NA7230, O404)	(NA5595, SA13)
(U100, NA7129)	(U80, O390)	(NA4756, O65)	(NA7355, O415)	(NA5687, SA18)
(U103, NA7199)	(U99, O402)	(NA4861, O83)	(NA7402, O419)	(NA6101, SA19)
(U104, NA7312)	(U118, O1032)	(NA5011, O105)	(NA7510, O422)	(NA6146, SA20)
(U106, NA7431)	(U181, O1104)	(NA5041, O114)	(NA7811, O625)	(NA6161, SA21)
(U107, NA7557)		(NA5188, O148)	(NA7832, O631)	(NA6185, SA22)
(U111, NA7679)	**Spike: Europe-South America**	(NA5194, O201)	(NA7845, O633)	(NA6299, SA25)
(U112, NA7884)	(U9, SA1)	(NA5200, O225)	(NA7901, O645)	(NA6373, SA27)

(U113, NA7914)	(U41, SA11)	(NA5205, O238)	(NA8514, O751)	(NA6395, SA28)
(U114, NA9075)	(U63, SA13)	(NA5372, O368)	(NA8646, O770)	(NA6396, SA29)
(U116, NA9180)	(U80, SA32)	(NA5538, O370)	(NA8703, O798)	(NA6406, SA30)
(U117, NA9189)	(U84, SA35)	(NA5579, O374)	(NA8787, O850)	(NA6418, SA31)
(U119, NA9264)	(U99, SA42)	(NA5595, O377)	(NA8817, O886)	(NA6429, SA32)
(U121, NA9283)	(U124, SA63)	(NA5819, O387)	(NA8824, O889)	(NA6515, SA33)
(U122, NA9284)	(U181, SA71)	(NA6334, O388)	(NA9091, O1017)	(NA6622, SA35)
(U123, NA9330)		(NA6395, O389)	(NA9333, O1035)	(NA6696, SA38)
(U126, NA9458)		(NA6429, O390)	(NA9350, O1037)	(NA7015, SA41)
(U131, NA10312)		(NA6577, O391)	(NA9639, O1059)	(NA7090, SA42)
(U137, NA10457)		(NA6578, O392)	(NA9792, O1076)	(NA7430, SA43)
(U141, NA10669)		(NA6620, O395)	(NA9891, O1079)	(NA7477, SA44)
(U144, NA10811)			(NA13635, O1104)	(NA7521, SA45)
(U146, NA10987)				(NA7892, SA56)
(U148, NA11013)				(NA9324, SA61)
(U151, NA11225)				(NA9910, SA66)
(U153, NA11367)				(NA10342, SA68)
(U154, NA11456)				(NA13390, SA70)
(U155, NA11466)				(NA13635, SA71)
(U158, NA13110)				
(U160, NA13253)				
(U175, NA13414)				
(U177, NA13551)				
(U179, NA13626)				
(U181, NA13635)				
(U187, NA14000)				

Table 12: List of spike proteins from Asia, which were found to be identical with spike proteins from other five continents

A1	A71	A115	A171	A207	A239	A280	A344	A388
A8	A76	A121	A173	A210	A244	A282	A345	A391
A12	A77	A122	A174	A211	A245	A283	A348	A394
A14	A78	A126	A175	A212	A247	A284	A351	A395
A15	A85	A127	A177	A213	A249	A286	A354	A396
A19	A89	A128	A178	A214	A253	A291	A356	A399
A23	A90	A133	A180	A215	A254	A292	A357	A401
A26	A91	A134	A181	A216	A255	A293	A358	A404
A28	A93	A135	A182	A217	A257	A304	A359	A405
A30	A95	A138	A183	A218	A258	A305	A360	A408
A31	A96	A140	A191	A219	A264	A322	A367	A413
A32	A97	A142	A193	A221	A265	A323	A373	A418
A33	A100	A148	A195	A222	A267	A324	A375	A419
A34	A101	A158	A196	A223	A273	A325	A378	A430
A36	A102	A159	A198	A224	A274	A333	A379	A431
A43	A103	A161	A199	A230	A275	A335	A380	
A44	A105	A163	A200	A233	A276	A340	A381	
A58	A109	A164	A201	A234	A277	A341	A383	
A67	A111	A166	A205	A235	A278	A342	A386	

A69	A114	A170	A206	A238	A279	A343	A387

Spike proteins (Asia) which were found to be identical with spike proteins from other five continents

Table 13: List of spike proteins from Africa, which were found to be identical with spike proteins from other five continents

AF1	AF34	AF58	AF79	AF101	AF117	AF128	AF145	AF156	AF227	AF263
AF2	AF38	AF64	AF82	AF103	AF118	AF130	AF146	AF165	AF229	AF268
AF3	AF46	AF69	AF83	AF104	AF119	AF131	AF147	AF168	AF230	AF271
AF8	AF47	AF71	AF88	AF105	AF120	AF133	AF149	AF179	AF231	AF278
AF9	AF48	AF72	AF90	AF108	AF121	AF134	AF151	AF195	AF247	AF283
AF19	AF50	AF73	AF92	AF114	AF123	AF137	AF152	AF196	AF248	
AF31	AF51	AF76	AF99	AF115	AF125	AF138	AF154	AF223	AF254	

Spike proteins (Afria) which were found to be identical with spike proteins from other five continents

Table 14: List of spike proteins from Europe, which were found to be identical with spike proteins from other five continents

U2	U18	U41	U63	U85	U103	U117	U137	U158
U3	U22	U42	U66	U88	U104	U118	U141	U160
U4	U23	U45	U72	U90	U106	U119	U144	U175
U7	U26	U47	U76	U92	U107	U121	U146	U177
U8	U28	U52	U78	U93	U111	U122	U148	U179
U9	U30	U53	U79	U94	U112	U123	U151	U181
U10	U34	U54	U80	U98	U113	U124	U153	U187
U11	U36	U55	U82	U99	U114	U126	U154	
U17	U38	U57	U84	U100	U116	U131	U155	

Spike proteins (Europe) which were found to be identical with spike proteins from other five continents

Table 15: List of spike proteins from North America, which were found to be identical with spike proteins from other five continents

NA7	NA3911	NA4837	NA5595	NA6161	NA6510	NA6810	NA7300	NA8703	NA9792	NA13390
NA231	NA3986	NA4861	NA56[illegible]6	NA6178	NA6515	NA6816	NA7312	NA8787	NA9834	NA13404
NA377	NA3988	NA4897	NA5627	NA6185	NA6527	NA6848	NA7355	NA8817	NA9891	NA13414
NA389	NA4024	NA4989	NA5644	NA6193	NA6540	NA6857	NA7375	NA8824	NA9910	NA13438
NA390	NA4028	NA5[illegible]1	NA5645	NA6240	NA6550	NA6862	NA7402	NA9075	NA10257	NA13444
NA402	NA4051	NA5011	NA5666	NA6244	NA6553	NA6903	NA7430	NA9091	NA10276	NA13465
NA902	NA4061	NA5022	NA5687	NA6258	NA6566	NA6916	NA7431	NA9180	NA10312	NA13478
NA928	NA4117	NA5041	NA5693	NA6276	NA6577	NA6936	NA7453	NA9189	NA10342	NA13551
NA992	NA4169	NA5083	NA5722	NA6293	NA6578	NA6944	NA7477	NA9264	NA10442	NA13554
NA1104	NA4243	NA5105	NA5744	NA6299	NA6602	NA6949	NA7510	NA9265	NA10457	NA13578
NA1131	NA4270	NA5137	NA5818	NA6305	NA6616	NA6953	NA7521	NA9283	NA10669	NA13614
NA1221	NA4296	NA5149	NA5819	NA6324	NA6620	NA6962	NA7553	NA9284	NA10811	NA13626
NA1298	NA4375	NA5151	NA5829	NA6334	NA6622	NA6969	NA7557	NA9324	NA10943	NA13635
NA1348	NA4394	NA5167	NA5830	NA6373	NA6628	NA7000	NA7576	NA9330	NA10987	NA13668
NA1445	NA4436	NA5182	NA5837	NA6388	NA6630	NA7015	NA7644	NA9333	NA11013	NA13704
NA2065	NA4448	NA5188	NA5874	NA6390	NA6659	NA7025	NA7679	NA9350	NA11135	NA13798
NA2680	NA4508	NA5194	NA5927	NA6395	NA6661	NA7056	NA7811	NA9425	NA11225	NA13841
NA322	NA453	NA520	NA597	NA639	NA668	NA705	NA783	NA945	NA1130	NA1387

8	7	0	7	6	3	7	2	5	5	0
NA3313	NA4541	NA5202	NA5992	NA6399	NA6687	NA7060	NA7845	NA9458	NA11367	NA13913
NA3387	NA4550	NA5205	NA6060	NA6406	NA6696	NA7090	NA7884	NA9568	NA11456	NA13948
NA3438	NA4559	NA5282	NA6067	NA6418	NA6704	NA7129	NA7892	NA9592	NA11466	NA14000
NA3477	NA4620	NA5372	NA6071	NA6421	NA6707	NA7198	NA7901	NA9597	NA11560	NA14015
NA3583	NA4637	NA5471	NA6080	NA6424	NA6713	NA7199	NA7914	NA9612	NA11576	NA14026
NA3658	NA4658	NA5485	NA6096	NA6429	NA6723	NA7224	NA8509	NA9639	NA11874	
NA3752	NA4715	NA5490	NA6101	NA6445	NA6751	NA7227	NA8514	NA9663	NA13110	
NA3768	NA4720	NA5516	NA6142	NA6450	NA6756	NA7230	NA8519	NA9674	NA13253	
NA3797	NA4756	NA5538	NA6146	NA6465	NA6775	NA7249	NA8565	NA9724	NA13280	
NA3873	NA4780	NA5574	NA6148	NA6492	NA6780	NA7286	NA8570	NA9763	NA13307	
NA3895	NA4832	NA5579	NA6155	NA6499	NA6794	NA7299	NA8646	NA9776	NA13362	

Spike proteins (North America) which were found to be identical with spike proteins from other five continents

Table 16: List of spike proteins from Oceania, which were found to be identical with spike proteins from other five continents

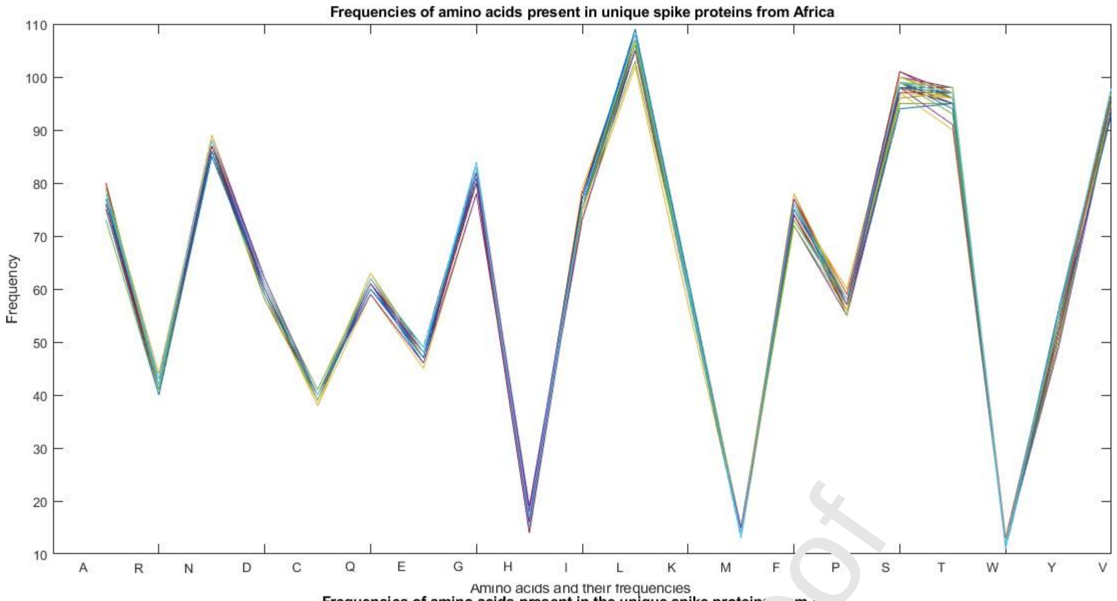
O3	O105	O373	O392	O419	O770	O1037
O5	O114	O374	O395	O422	O798	O1039
O28	O148	O377	O398	O504	O850	O1076
O36	O201	O387	O400	O625	O886	O1079
O43	O225	O388	O401	O631	O889	O1104
O58	O238	O389	O402	O633	O1017	
O65	O368	O390	O404	O645	O1032	
O83	O370	O391	O415	O751	O1035	

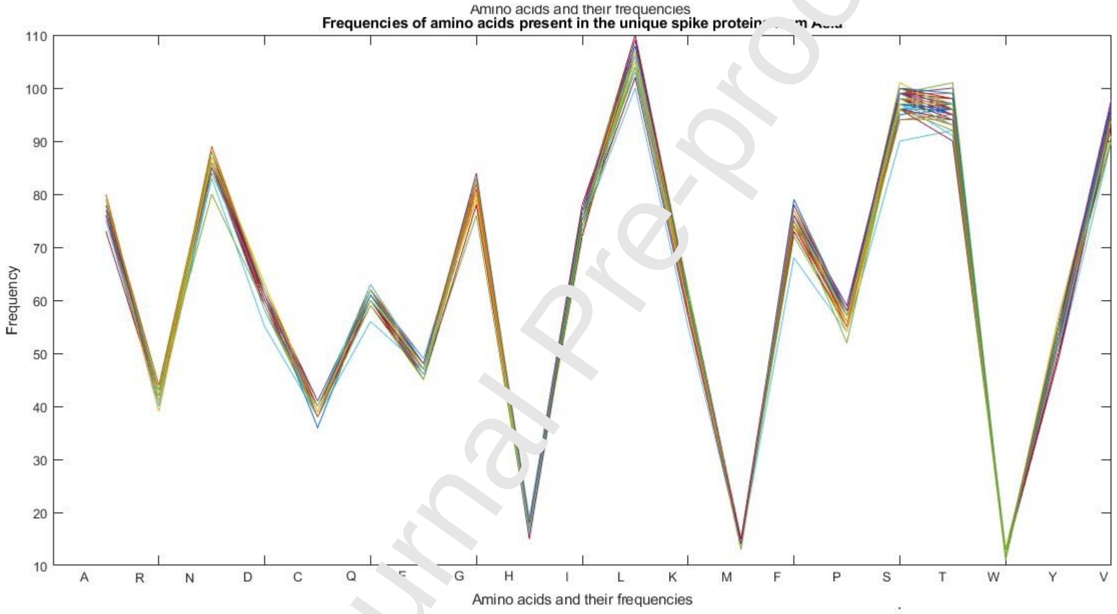
Spike proteins (Oceania) which were found to be identical with spike proteins from other five continents

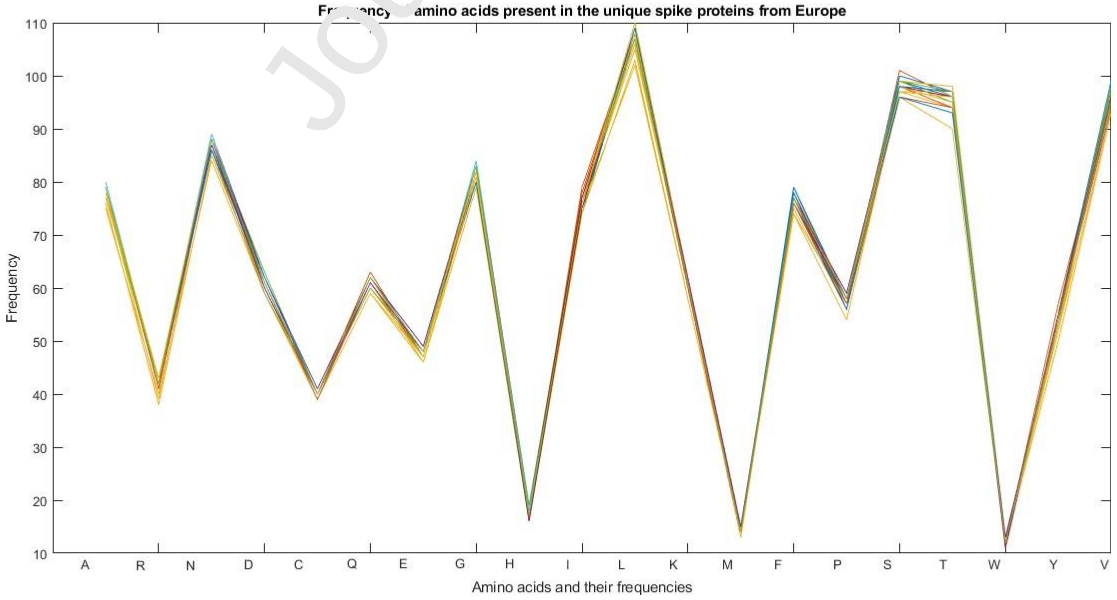
Table 17: List of spike proteins from South America, which were found to be identical with spike proteins from other five continents

SA1	SA13	SA22	SA29	SA35	SA44	SA66
SA4	SA18	SA25	SA30	SA38	SA45	SA68
SA5	SA19	SA26	SA31	SA41	SA56	SA70
SA7	SA20	SA27	SA32	SA42	SA61	SA71
SA11	SA21	SA28	SA33	SA43	SA63	

Spike proteins (Oceania) which were found to be identical with spike proteins from other five continents


Frequencies of amino acids present in unique spike proteins from Africa
Frequency
110
100
90
80
70
60
50
40
30
20
10
A
R
N
D
C
Q
E
G
H
I
L
K
M
F
P
S
T
W
Y
V
Amino acids and their frequencies


Frequencies of amino acids present in the unique spike proteins from Asia
Frequency
110
100
90
80
70
60
50
40
30
20
10
A
R
N
D
C
Q
E
G
H
I
L
K
M
F
P
S
T
W
Y
V
Amino acids and their frequencies


Frequency of amino acids present in the unique spike proteins from Europe
Frequency
110
100
90
80
70
60
50
40
30
20
10
A
R
N
D
C
Q
E
G
H
I
L
K
M
F
P
S
T
W
Y
V
Amino acids and their frequencies

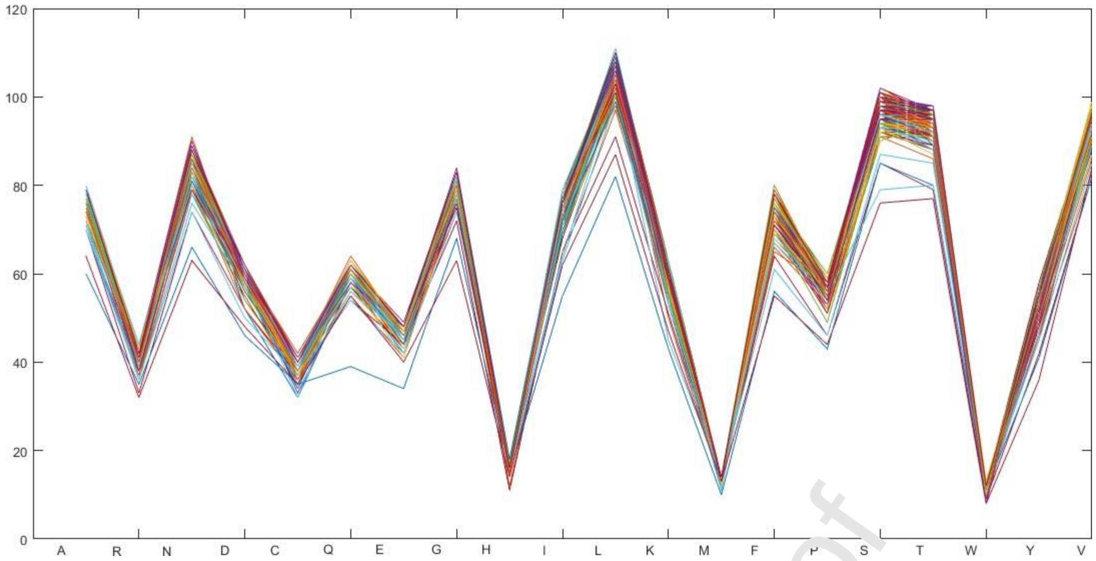
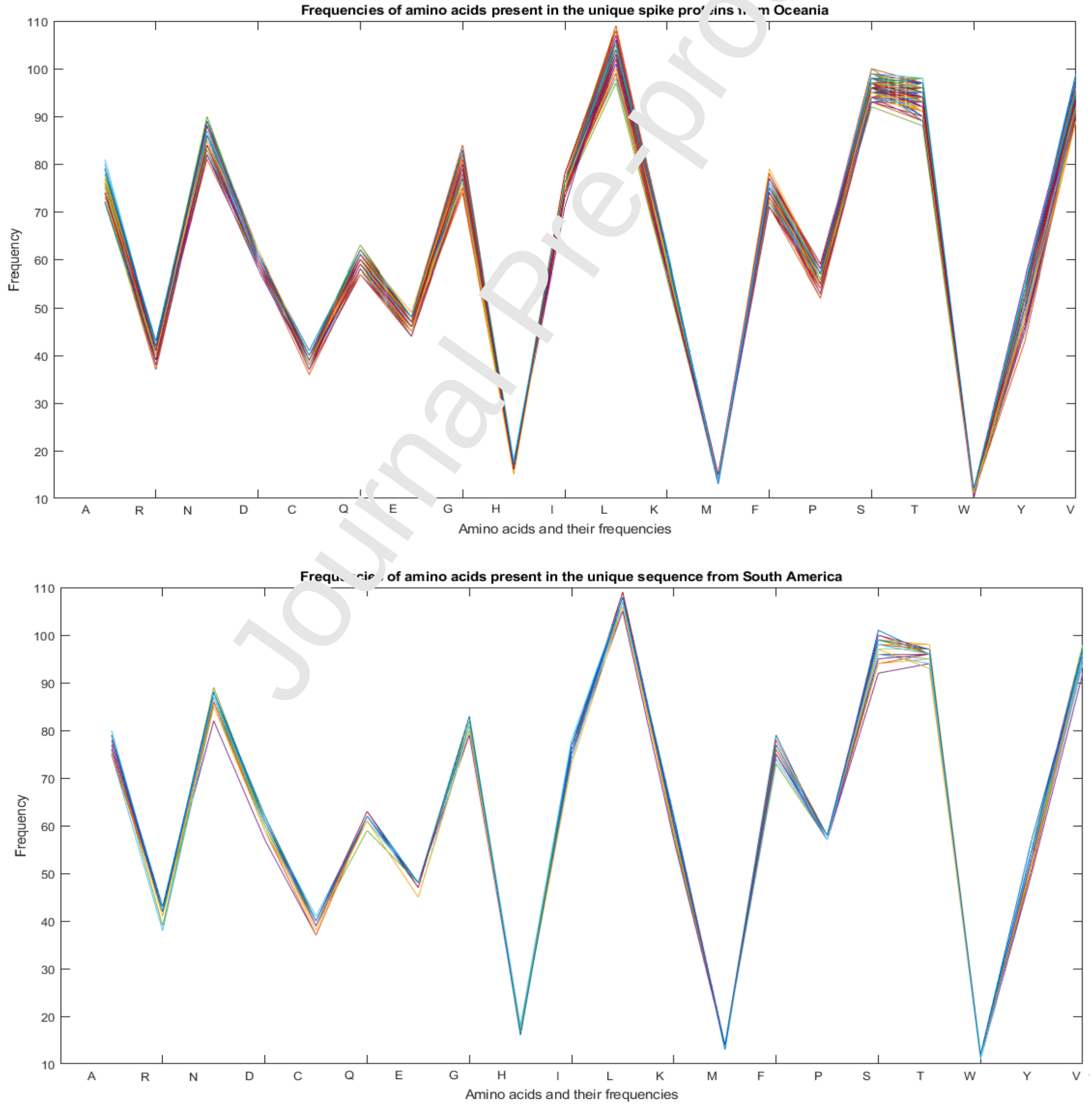
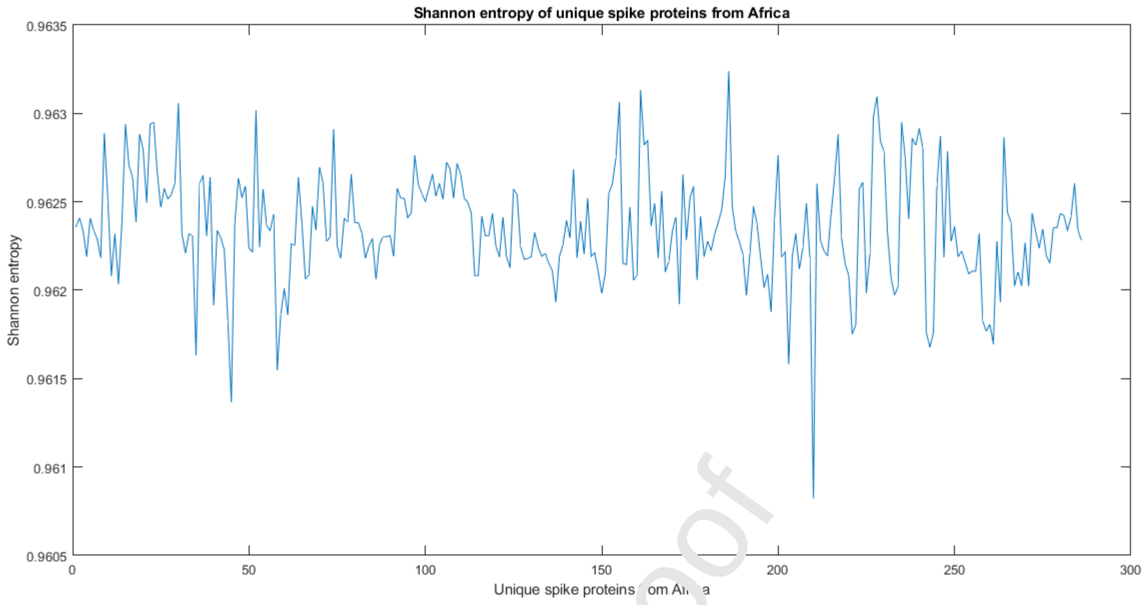


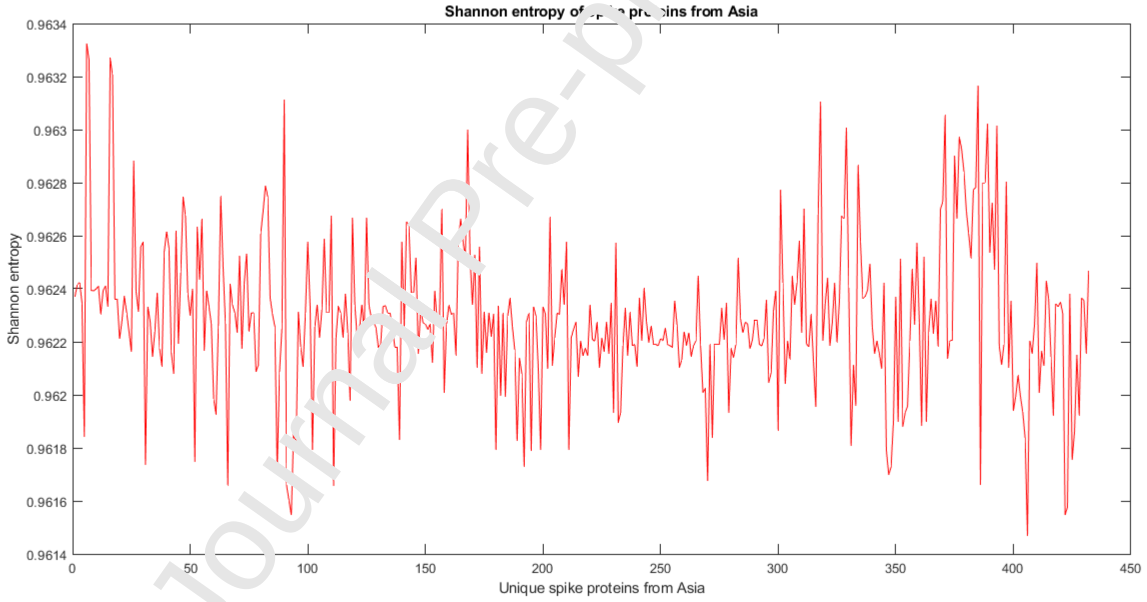
Figure 6: Frequencies of amino acids present in the unique S sequences


Frequencies of amino acids present in the unique spike proteins from Oceania

Frequencies of amino acids present in the unique sequence from South America

Figure 7: Frequencies of amino acids present in the unique S sequences


Shannon entropy of unique spike proteins from Africa
Shannon entropy
0.9635
0.963
0.9625
0.962
0.9615
0.961
0.9605
0
50
100
150
200
250
300
Unique spike proteins from Africa


Shannon entropy of spike proteins from Asia
Shannon entropy
0.9634
0.9632
0.963
0.9628
0.9626
0.9624
0.9622
0.962
0.9618
0.9616
0.9614
0
50
100
150
200
250
300
350
400
450
Unique spike proteins from Asia

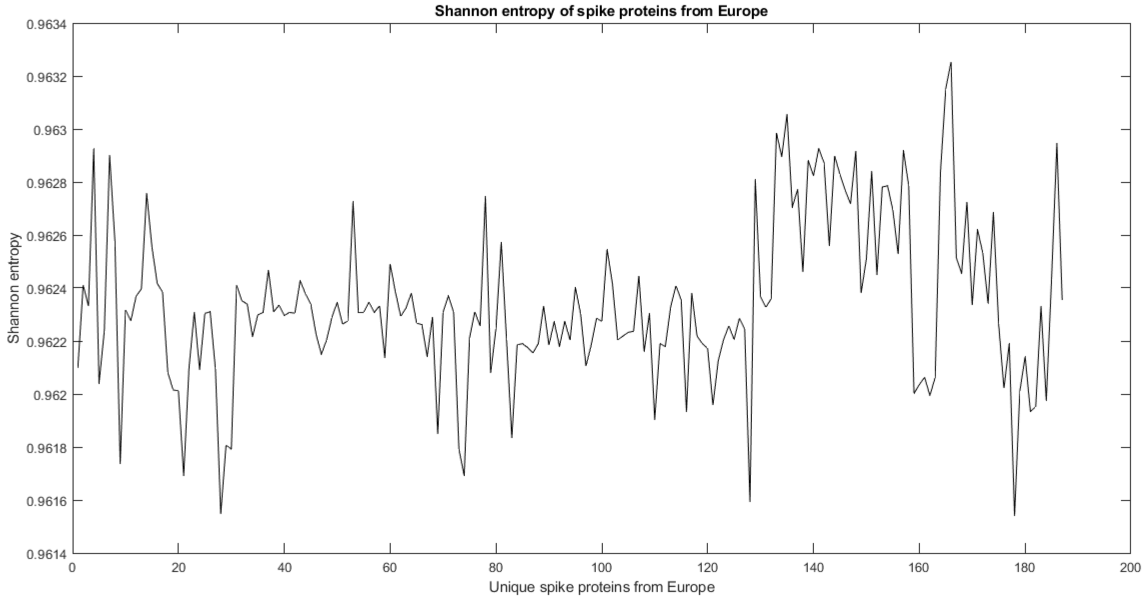

Shannon entropy of spike proteins from Europe
Shannon entropy
0.9634
0.9632
0.963
0.9628
0.9626
0.9624
0.9622
0.962
0.9618
0.9616
0.9614
0
20
40
60
80
100
120
140
160
180
200
Unique spike proteins from Europe

Figure 8: SE of unique S proteins from different continents

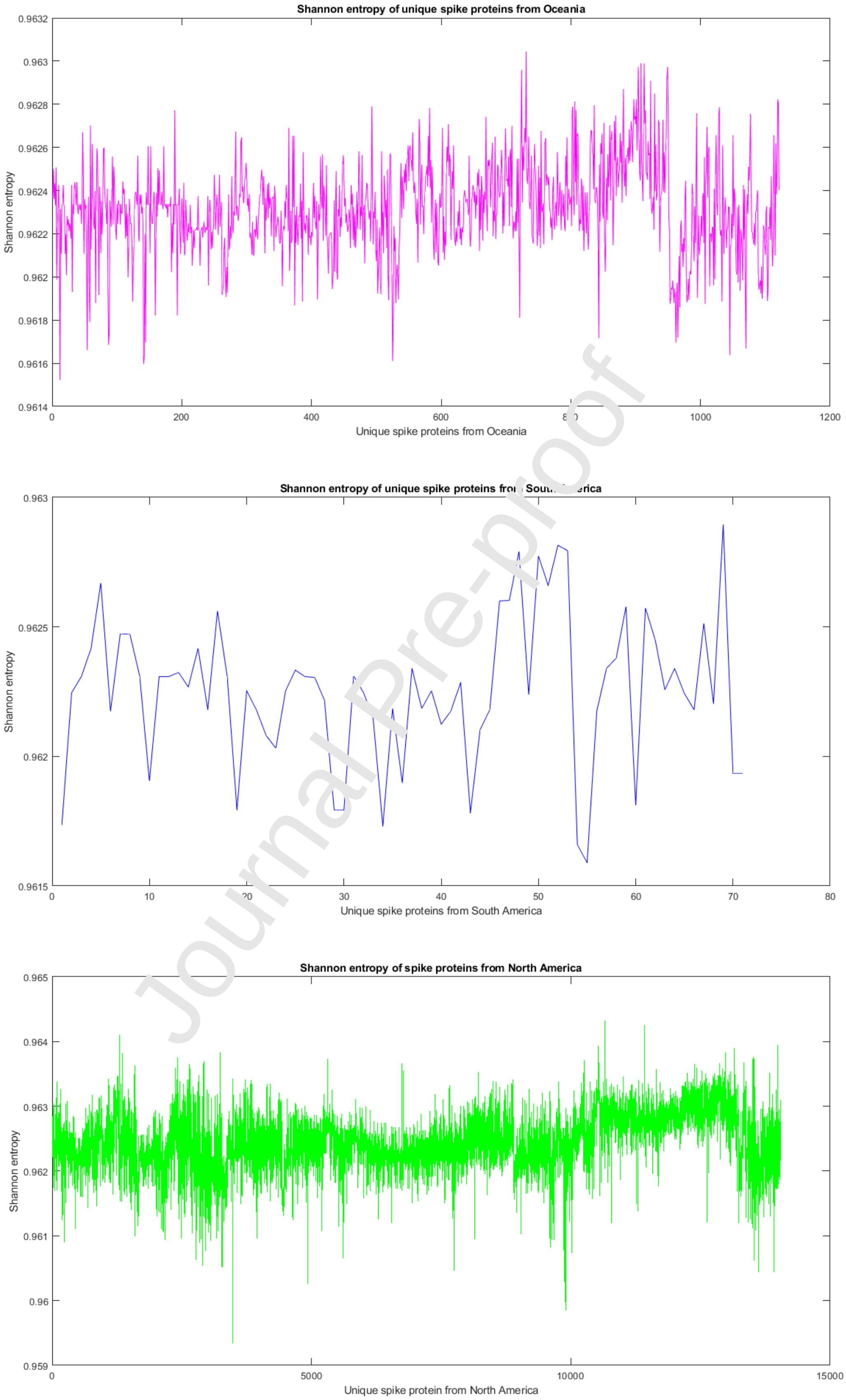


Figure 9: SE of unique S proteins from different continents

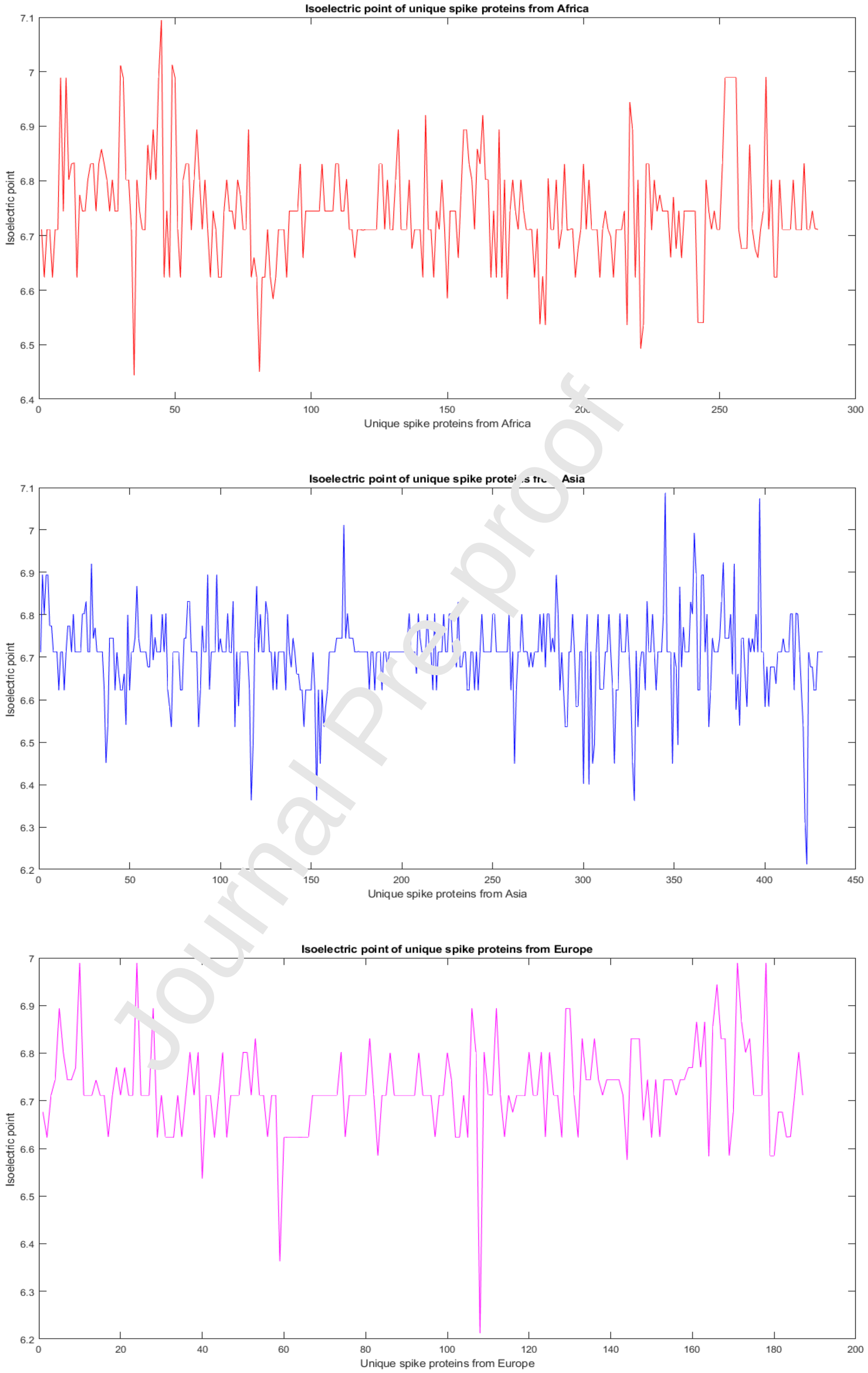


Figure 10: Isoelectric point of unique S proteins from different continents

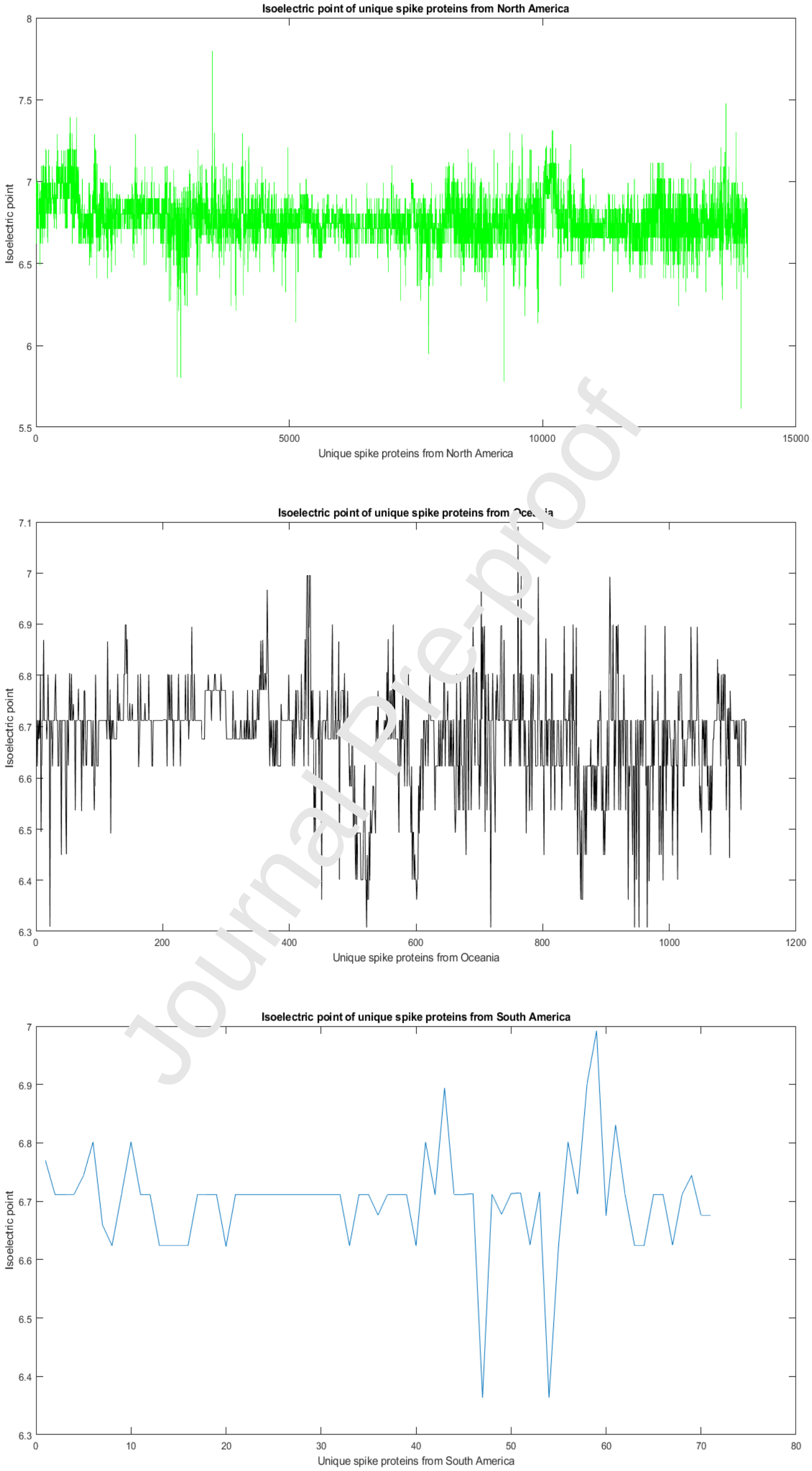


Figure 11: Isoelectric point of unique S proteins from different continents